

VECTORS FOR DIRECTIONAL CLONING

Background of the Invention

5 Molecular biotechnology has revolutionized the production of protein compounds of pharmacological importance. The advent of recombinant DNA technology permitted for the first time the production of proteins on a large scale in a recombinant host cell rather than by the laborious and expensive isolation of the protein from cells or tissues which may contain minute quantities of that protein. The production of proteins, including human proteins, on a large scale in a host
10 requires the ability to express the protein of interest in a host cell, e.g., a heterologous host cell. This process typically involves isolation or cloning of the gene encoding the protein of interest followed by transfer of the coding region (open reading frame) into an expression vector which contains elements (e.g., promoters) which direct the expression of the desired protein in the host cell. The most
15 commonly used means of transferring or subcloning a coding region into an expression vector involves the *in vitro* use of restriction endonucleases and DNA ligases. Restriction endonucleases are enzymes which generally recognize and cleave a specific DNA sequence in a double-strand DNA molecule. Restriction enzymes are used to excise a DNA fragment which includes a coding region of
20 interest from the cloning vector and the excised DNA fragment is then joined using DNA ligase to a suitably cleaved vector with transcription regulatory sequences in such a manner that a functional protein can be expressed when the resulting expression vector is introduced to a cell or an *in vitro* transcription/translation mixture.

25 A problem in controlling fragment orientation in fragments generated by restriction enzymes is that many of the commonly used restriction enzymes produce termini that are rotationally equivalent, and therefore, self-ligation of DNA fragments with such termini is random with regard to fragment orientation. Hartley and Gregori (Gene, 13:347 (1981)) reported a technique to control fragment
30 orientation during ligation, which required the introduction of *Ava*I sites flanking either end of the cloned fragment (also see Hartley and Gregori, U.S. Patent No.

4,403,036). Since *Ava*I cleavage produces distinguishable ends, self-ligation of the fragment results in a strong bias toward head-to-tail orientation. This is so because head-to-head and tail-to-tail ligation results in base mismatches. The polymerized molecules were then inserted into a vector and used to transform *E. coli*.

5 In a similar approach, Ikeda et al. (Gene, 71:19 (1988)) produced head-to-tail tandem arrays of a DNA fragment encoding a human major histocompatibility antigen that was flanked by *Sfi*I cleavage sites. *Sfi*I produces single-strand DNA overhangs that are not rotationally equivalent. *Sfi*I sites have also been used to produce copolymers of gene expression cassettes and selection markers, which can
10 be used to transfect cells (Monaco et al., Biotechnol. Appl. Biochem., 20:157 (1994); Asselbergs et al., Anal. Biochem., 243:285 (1996)). Monaco et al. treated the copolymer with *Not*I to cleave the DNA at the 3' end of the selectable marker gene. In this way, transfected DNA molecules contain only one selectable marker gene per copolymer.

15 Class IIS restriction enzymes can generate totally asymmetric sites and complementary cohesive ends. Kim and Szybalski (Gene, 71:1 (1988)) introduced sites for *Bsp*MI, a class IIS restriction enzyme, at either end of cloned DNA. Self-ligation of the cloned DNA provided multimers comprising repeat units in the same orientation. Similarly, Takeshita et al. (Gene, 71:9 (1988)) achieved tandem gene
20 amplification by inserting a fragment encoding human protein C into a plasmid to introduce asymmetric cohesive ends into the fragment. In this case, sites for the class IIS enzyme, *Bst*XI, were used. The multimer was then cloned into a cosmid vector comprising a *neo* gene, packaged into lambda phage particles, and amplified in *E. coli*. The cosmid vectors were then introduced into Chinese hamster ovary
25 DHFR-cells, which were treated with G418 to select for cells that expressed the *neo* gene. Takeshita et al. also found that cells expressed human protein C, albeit at lower levels, following transfection with unpackaged tandem ligated DNA comprising copies of the cosmid vector and the human protein C gene.

 A similar approach was described by Lee et al. (Genetic Analysis: Biomolecular Engineering, 13:139 (1996)), who amplified target DNA as tandem
30 multimers by cloning the target DNA into a class IIS restriction enzyme cleavage

site of a vector, excising a monomeric insert with the class IIS restriction enzyme, isolating monomeric inserts, self-ligating the inserts, and cloning the multimers into a vector. According to Lee et al., such a method is useful for polymerizing short DNA fragments for the mass production of peptides.

5 Another approach for forcing directional ligation is to devise synthetic linkers or adapters that are used to create asymmetric cohesive ends. For example, Taylor and Hagerman (Gene, 53:139 (1987)) modified the Hartley-Gregori approach by attaching synthetic directional adapters to a DNA fragment in order to establish control over fragment orientation during ligation. Following
10 polymerization, the multimers were ligated to a linearized vector suitable for *E. coli* transformation. Stahl et al. (Gene, 89:187 (1990)) described a similar method for polymerizing DNA fragments in a head-to-tail arrangement. Here, synthetic oligonucleotides were designed to encode an epitope-bearing peptide with 5'-protruding ends complementary to the asymmetric cleavage site of the class IIS
15 restriction enzyme, *Bsp*MI. After polymerization, the peptide encoding fragments were inserted into the unique *Bsp*MI site cleavage site of a vector, which was used to transform *E. coli*. Clones were screened using the polymerase chain reaction, and then subcloned into prokaryotic expression vectors for production of the peptides in *E. coli*.

20 Nevertheless, the ability to transfer a desired coding region to a vector with transcription regulatory sequences is often limited by the availability or suitability of restriction enzyme recognition sites. Often multiple restriction enzymes must be employed for the removal of the desired coding region and the reaction conditions used for each enzyme may differ such that it is necessary to perform the excision
25 reactions in separate steps. In addition, it may be necessary to remove a particular enzyme used in an initial restriction enzyme reaction prior to completing remaining restriction enzyme digestions. This requires a time-consuming purification of the subcloning intermediate. It also may be necessary to inactivate restriction enzymes prior to ligation.

30 Methods for the directional transfer of a target DNA molecule from one vector to another *in vitro* or *in vivo* without the need to rely upon restriction enzyme

digestions have been described. For example, the Creator™ DNA cloning kit (Clontech Laboratories, Inc.) uses Cre-*loxP* site-specific recombination to catalyze the transfer of a target gene from a donor vector to an acceptor vector, which is a plasmid containing regulatory elements of the desired host expression system (see also U.S. Patent No. 5,851,808). Cre, a 38-kDa recombinase protein from bacteriophage P1, mediates recombination between or within DNA sequences at specific locations called *loxP* sites (Sauer, *Biotechniques*, 16:1086 (1994); Abremski et al., *J. Biol. Chem.*, 259:1509 (1984)). These sites consist of two 13 bp inverted repeats separated by an 8 bp spacer region that provides directionality to the recombination reaction. The 8 bp spacer region in the *loxP* site has a defined orientation which forces the target gene to be transferred in a fixed orientation and reading frame. Donor vectors in the kit contain two *loxP* sites, which flank the 5' end of a multiple cloning site (MCS) and the 5' end of the open reading frame for the chloramphenicol resistance gene. Donor vectors also contain the ampicillin gene for propagation and selection in *E. coli*, and the sucrose gene from *B. subtilis* (*SacB*) for selection of correct recombinants. Acceptor vectors in the kit contain a single *loxP* site, followed by a bacterial promoter, which drives expression of the chloramphenicol marker after Cre-*lox*-mediated recombination. The gene of interest, once transferred, becomes linked to the specific expression elements for which the acceptor vector was designed. If the coding sequence for the gene of interest is in frame with the upstream *loxP* site in the donor vector, it is in frame with all peptides in the acceptor vector.

The Gateway™ Cloning System uses phage lambda-based site-specific recombination. The LR Reaction is a recombination reaction between an entry clone having mutant *attL* sites and a vector (a Destination Vector, pDEST™) having the corresponding mutant *attR* site, mediated by a cocktail of recombination proteins (λ recombination proteins Int, Xis, and the *E. coli*-encoded protein IHF), to create an expression clone. The BP Reaction is a recombination reaction between an expression clone (or an *attB*-flanked PCR product) and a donor vector to create an entry clone. The BP reaction permits rapid, directional cloning of PCR products synthesized with primers containing terminal 25 bp *attB* sites (+4 Gs). The result is

an entry clone containing the PCR fragment. Similarly, DNA segments flanked by *attB* sites in an expression clone can be transferred to generate entry clones which can be used to move the sequence of interest to one or more destination vectors in parallel reactions to generate expression clones. The resultant 25 bp *attB* sites (*attB*1 on the left (N-terminus) and *attB*2 on the right (C-terminus)) created by the LR reaction are derived from the *attL* sites (adjacent to the gene), whereas the distal sequences are derived from the *attR* sites.

However, the protein encoded by Cre-*loxP* based expression vectors or other site-specific recombinase based vectors, e.g., the Gateway™ Cloning System, has numerous, for instance, 8 to 13, amino acid residues at the N-terminus and C-terminus of the protein, which residues are encoded by the site-specific recombination exchange sites.

Thus, what is needed is an improved method to directionally clone a nucleic acid sequence of interest.

Summary of the Invention

The invention provides methods and vectors for use in directional cloning. In one embodiment, a vector comprising an open reading frame of interest (a donor vector) comprises at least two restriction enzyme recognition sites (“restriction enzyme sites”, “restriction sites” or “recognition sites”) flanking the open reading frame (DNA sequence of interest), wherein at least one of the flanking sites is a site for a first restriction enzyme which generates hapaxotermistic ends, e.g., a restriction enzyme with a degenerate recognition sequence or one which cleaves outside of a recognition sequence yielding single-strand ends, and other vector sequences (backbone sequences) for replication and/or maintenance of the vector in a host cell and, optionally, one or more detectable, e.g., selectable, marker genes. In one embodiment, a donor vector comprises at least two restriction enzyme sites flanking the open reading frame, wherein at least one of the flanking sites is for a first restriction enzyme which is a hapaxotermistic restriction enzyme, e.g., a restriction enzyme with a degenerate recognition sequence, which site, once cleaved, does not result in self complementary single-strand DNA overhangs or

blunt ends, i.e., the ends are non-self complementary single-strand DNA overhangs. In another embodiment, the donor vector comprises at least two restriction enzyme sites flanking the open reading frame, wherein at least two of the flanking sites are for a first restriction enzyme with a hapaxomeric recognition sequence, and optionally for the same restriction enzyme, which sites, once cleaved, yield a linear DNA fragment which does not have self-complementary single-strand DNA overhangs or blunt ends. Such a vector may be employed as a source of the open reading frame to prepare a vector for expression of the linked open reading frame (a recipient or expression vector). The backbone sequences in the recipient vector are generally provided by an acceptor vector which contains transcriptional regulatory sequences and optionally sequences for the production of fusion proteins. The acceptor vector also comprises non-essential DNA sequences flanked by at least two restriction enzyme sites for a second restriction enzyme with a hapaxomeric recognition sequence, and optionally one or more detectable, e.g., selectable, marker genes. In one embodiment, the two flanking restriction enzyme sites in the acceptor vector for the second restriction enzyme are sites which, once cleaved, do not result in self complementary single-strand DNA overhangs or blunt ends but yield a linear DNA fragment having single-strand DNA overhangs that are complementary with one of the two DNA overhangs generated by the first restriction enzyme. Once the linearized DNA fragments are ligated to form a recipient vector, the recipient vector may be introduced to cells, e.g., prokaryotic cells such as *E. coli* cells, insect cells, plant cells, mammalian cells, or lysates thereof or to *in vitro* transcription/translation mixtures, so as to yield a transformed cell that expresses a protein encoded at least in part by the open reading frame.

In one embodiment, the invention provides a method for the directional subcloning of DNA fragments. The method includes providing a first vector comprising a first selectable marker gene and a DNA sequence of interest, which DNA sequence of interest is flanked by at least two restriction enzymes sites, wherein at least two of the flanking restriction enzyme sites are sites for a first restriction enzyme which is a hapaxotermistic restriction enzyme, and wherein digestion of the first vector with the first restriction enzyme generates a first linear

DNA fragment which lacks the first selectable marker gene but comprises the DNA sequence of interest and a first pair of non-self complementary single-strand DNA overhangs. A second vector for the method is provided which includes a second selectable marker gene which is distinguishable from the first selectable marker gene and non-essential DNA sequences, optionally including a counterselectable gene, which non-essential DNA sequences are flanked by at least two restriction enzyme sites, wherein at least two of the flanking restriction enzyme sites are for a second restriction enzyme which is a hapaxotermistic restriction enzyme, wherein digestion of the second vector with the second restriction enzyme generates a second linear DNA fragment which lacks the non-essential DNA sequences but comprises the second selectable marker gene and a second pair of non-self complementary single-stranded DNA overhangs, and wherein each of the second pair of non-self complementary single-strand DNA overhang is complementary to only one of the single-strand DNA overhangs of the first pair of non-self complementary single-strand DNA overhangs and permits the oriented joining of the first linear DNA fragment to the second linear DNA fragment. The first and second vectors, the first vector and the second linear DNA fragment, or the second vector and the first linear DNA fragment are combined in a suitable buffer with one or more of the restriction enzymes which are hapaxotermistic restriction enzymes and optionally DNA ligase under conditions effective to result in digestion and optionally ligation to yield a mixture optionally comprising a third vector comprising the first and second linear DNA molecules which are joined in an oriented manner via the first and second pairs of non-self-complementary single-strand DNA overhangs. In one embodiment, ligase is added simultaneously with the one or more restriction enzymes, while in another embodiment, ligase is added subsequent to the one or more restriction enzymes. Optionally, the mixture is introduced into a host cell, and optionally the transformed host cells are selected for the expression of second selectable marker gene or against the expression of the counterselectable gene. The method may also include identifying a third vector in which the DNA sequence of interest has been transferred in an oriented manner to the second linear DNA fragment. In one embodiment, the first restriction enzyme is

SfiI, *SapI* or an isoschizomer thereof. In one embodiment, the first restriction enzyme is *SfiI* or an isoschizomer thereof and the second restriction enzyme is *BglI* or an isoschizomer thereof. In one embodiment, the second restriction enzyme is *EaeI* or an isoschizomer thereof. In another embodiment, the first and second

5 restriction enzymes are the same. Optionally, the DNA sequence of interest comprises an open reading frame comprising one or more sites for the first or second restriction enzyme. In this embodiment, optionally, prior to digestion with the one or more restriction enzymes, the sites for the one or more restriction enzymes in the open reading frame are protected so as to prevent digestion, e.g.,

10 protected by methylation such as with *HaeIII* methylase, *SapI* methylase, or *SfiI* methylase. Alternatively, prior to methylation, the flanking sites for the first or second restriction enzyme are contacted with an oligonucleotide complementary to the flanking restriction enzyme site and RecA. In one embodiment, ligation and oriented joining yields a third vector encoding a N-terminal fusion protein which is

15 encoded by the DNA sequence of interest and nucleic acid sequences 5' to the 3' end of the second linear DNA fragment. In another embodiment, ligation and oriented joining yields a third vector encoding a C-terminal fusion protein which is encoded by the DNA sequence of interest and nucleic acid sequences 3' to the 5' end of the second linear DNA fragment. In yet another embodiment, ligation and oriented

20 joining yields a third vector encoding a fusion protein which is encoded by the DNA sequence of interest and nucleic acid sequences 5' and 3' to the respective 3' and 5' end of the second linear DNA fragment. In a further embodiment, ligation and oriented joining yields a third vector encoding a fusion protein which is encoded by the DNA sequence of interest and the exchange site(s) created by the oriented

25 joining.

Thus, the invention also provides a vector system for cloning. In one embodiment the system includes a first vector comprising a selectable marker gene and a DNA sequence of interest, which DNA sequence of interest is flanked by at least two restriction enzyme sites, wherein at least two of the flanking restriction

30 enzyme sites are for a first restriction enzyme which is a hapaxotermistic restriction enzyme, wherein digestion of the first vector with the first restriction

enzyme generates a first linear DNA fragment which does not comprise the first selectable marker gene but comprises the DNA sequence of interest and a first pair of non-self complementary single-strand DNA overhangs, wherein the first restriction enzyme sites are designed such that the first linear DNA fragment can be
5 religated directly to a second vector. The system optionally includes a second vector, which includes a second selectable marker gene which is distinguishable from the first selectable marker and non-essential DNA sequences, optionally including a counterselectable gene, which non-essential DNA sequences are flanked by at least two restriction enzyme sites, wherein two or more of the flanking
10 restriction enzyme sites in the second vector are for a second restriction enzyme which is a hapaxotermistic restriction enzyme, wherein digestion of the second vector with the second restriction enzyme generates a second linear DNA fragment which lacks the non-essential DNA sequences but comprises the second selectable marker gene and a second pair of non-self complementary single-strand DNA
15 overhangs, wherein each of the second pair of non-self complementary single-strand DNA overhangs is complementary to only one of the single-strand DNA overhangs of the first pair of non-self complementary single-strand DNA overhangs and permits the oriented joining of the first linear DNA fragment to the second linear DNA fragment. Further provided is a kit which includes one or more vectors of the
20 vector system.

In one embodiment, at least one restriction enzyme site flanking the open reading frame of interest is for a restriction enzyme that recognizes an internal palindrome, e.g., a type II enzyme such as *SfiI* or *BglII*, including but not limited to restriction enzymes that generate more than two types of staggered ends (DNA
25 overhangs) due to the ambiguity in base recognition, for instance, *AhdI*, *AlwNI*, *ApaBI*, *BglII*, *BlpI*, *BstAPI*, *BstEII*, *BstXI*, *Bsu36I*, *DraII*, *DraIII*, *DrdI*, *Eam1105I*, *EcoNI*, *PflMI*, *PssI*, *SauI*, *SfiI*, *XcmI*, as well as isoschizomers thereof, but not restriction enzymes that generate blunt ends. In another embodiment, at least one restriction enzyme site flanking the open reading frame of interest is for a type IIS
30 enzyme, e.g., *SapI* or *EarI*, such as restriction enzymes that generate ends outside of their recognition sites including but not limited to *AarI*, *AceIII*, *AloI*, *BaeI*, *Bbr7I*,

BbvI, BbvII, BccI, Bce83I, BceAI, BceFI, BcgI, BciVI, BfiI, BinI, BpII, BsaXI, BscAI, BseMII, BseRI, BsgI, BsmI, BsmAI, BsmFI, Bsp24I, BspCNI, BspMI, BsrI, BsrDI, BstF5I, BtgZI, BtsI, CjeI, CjePI, EciI, Eco31I, Eco57I, Eco57MI, Esp3I, 5 FalI, FauI, FokI, GsuI, HaeIV, HgaI, Hin4I, HphI, HpyAV, Ksp632I (EarI), MboII, MlyI, MmeI, MnlI, PleI, PpiI, PsrI, RleAI, SapI, VapK32I, SfaNI, SspD5I, Sth132I, StsI, TaqII, TspDTI, TspGWI, TspRI, Tth111II, as well as isoschizomers thereof. In one embodiment, the open reading frame encodes one or more domains of one or more proteins.

In another embodiment, the invention provides a donor vector comprising an
10 open reading frame of interest flanked by at least two restriction enzyme sites, one of which flanking sites is for a first restriction enzyme that has a low frequency, e.g., fewer than about 25%, for instance, including fewer than about 20%, 10%, 5% or even fewer, e.g., about 1%, of recognition sites in a plurality of, for instance, 3 or more, including 100, 1,000, 10,000 or more, cDNAs or open reading frames for a
15 particular species (an “infrequent cutter”) and generates single-strand DNA overhangs, and the other of which flanking sites is for a second restriction enzyme that has a low frequency of recognition sites in a plurality of cDNAs or open reading frames for a particular species, for instance, the same species as for the first restriction enzyme, and generates ends that are not complementary to the overhangs
20 generated by the first restriction enzyme. In one embodiment, the second restriction enzyme generates blunt ends (a “blunt cutter”). The frequency of a particular restriction enzyme recognition site in one or more nucleic acid molecules can be determined by methods well-known to the art. For instance, databases with a plurality of cDNA sequences or open reading frames for a particular organism may
25 be employed to determine such a frequency. A donor vector of the invention may be employed as a source of the open reading frame of interest to prepare a recipient vector of the invention. The backbone sequences in the recipient vector are generally provided by an acceptor vector having transcriptional regulatory sequences of interest and optionally sequences for the production of fusion proteins.
30 The acceptor vector also comprises non-essential DNA sequences flanked by at least two restriction enzyme sites, and one or more detectable marker genes. In one

embodiment, one of the flanking sites in the acceptor vector is for a third restriction enzyme which generates single-strand DNA overhangs, which single-strand DNA overhangs are complementary with the single-strand DNA overhangs produced when the donor vector is digested with the first restriction enzyme. The other

5 flanking site in the acceptor vector is for a fourth restriction enzyme which generates ends that are not complementary to the ends generated by the first or third restriction enzyme but are compatible, i.e., can be ligated to, with ends generated by the second restriction enzyme. In one embodiment, the second and fourth

10 recognition enzymes are blunt cutters and the restriction sites for the second and fourth restriction enzymes are not recognized by the same restriction enzyme. In one embodiment, at least one restriction enzyme site flanking the open reading frame of interest is for one of *SgfI*, *PvuI* or *PacI*, restriction enzymes which generate ends compatible with *SgfI*, e.g., *SgfI*, *PvuI*, *BstKTI* or *PacI*, or restriction enzymes that yield ends that can be selected to have the proper 3' TA overhang, e.g., *AasI*,

15 *Bce83I*, *BstEI*, *BcgI*, *BpmI*, *BpuEI*, *BseMI*, *Bse3DI*, *BseMII*, *BseRI*, *BsgI*, *BspCNI*, *BsrDI*, *BstF5I*, *BseGI*, *BtsI*, *DrdI*, *DseDI*, *EciI*, *Eco57MI*, *Eco57I*, *Bce83I*, *GsuI*, *MmeI*, *TspDTI*, *Tth111II*, *BspKT5I*, *AcuI*, *BspKT6I*, *Eco57MI*, *TaqII*, *TspGWI*, or isoschizomers thereof. In one embodiment, at least one restriction enzyme site flanking the open reading frame of interest is for one of *SgfI* (*AsiSI*), *PacI*, or *PvuI*

20 (*Afa22MI*, *Afa16RI*, *BspCI*, *EagBI*, *ErhB9I*, *MvrI*, *NbII*, *Ple19I*, *Psu161I*, *RshI*, *XorII*).

In another embodiment, at least one restriction enzyme site flanking the open reading frame of interest is for *PmeI* (*MssI*), *DraI*, *AhaIII* (*DraI*, *PauAII*, *SruI*), *NruI* (*Bsp68I*, *MluB2I*, *Sbo13I*, *SpoI*), *SnaBI* (*BstSNI*, *Eco105I*), *SrfI*, or *SwaI*

25 (*BstRZ246I*, *BstSWI*, *MspSWI*, *SmiI*). In another embodiment, at least one restriction enzyme site flanking the open reading frame of interest is for a restriction enzyme that generates a blunt end which can create a stop codon after ligation with another blunt end, for instance, one that can create a stop codon after ligation with an end generated by *PmeI*, e.g., *EcaBC3I* (TC[^]GA), *SciI* (CTC[^]GAG), *HincII* (GTC[^]GAC, a version of GTYRAC), *HpaI* (GTT[^]AAC), *HincII* (GTT[^]AAC, a

30 version of GTYRAC), *DraI* (TTT[^]AAA), *SwaI* (ATTT[^]AAAT), or an isoschizomer

thereof, or for a restriction enzyme that yields ends that can be selected to have a blunt end such as 5'GA, 5'AG or 5'AA, e.g., *BsaBI*, *Cac8I*, *Hpy8I*, *MlyI*, *PshAI*, *SspD5I*, or an isoschizomer thereof. For example, ligation of ends generated by *PmeI* and *DraI* can create a stop site, as would ligation of NTT and GAN, NCT and AGN, or NTT and AAN, wherein each N is A, T, G or C. In one embodiment, the exchange site formed from blunt end ligation of an end generated by *PmeI* and that of another blunt cutter can yield a coding sequence for a protein fusion. For instance, ligation of an open reading frame terminating in an end generated by *PmeI* and an end generated by *BalI*, *BfrBI*, *BsaAI*, *BsaBI*, *BsrBI*, *BtrI*, *Cac8I*, *CdiI*, *CviII*, *CviRI*, *Eco47III*, *Eco78I*, *EcoICRI*, *EcoRV*, *FnuDII*, *FspAI*, *HaeI*, *HaeIII*, *Hpy8I*, *LpnI*, *MlyI*, *MslI*, *MstI*, *NaeI*, *NlaIV*, *NruI*, *NspBII*, *OliI*, *PmaCI*, *PmeI*, *PshAI*, *PsiI*, *PvuII*, *RsaI*, *ScaI*, *SmaI*, *SnaBI*, *SrfI*, *SspI*, *SspD5I*, *StuI*, *XcaI*, *XmnI*, *ZraI*, or an isoschizomer thereof, can extend the open reading frame at the 3' end.

In one embodiment, the invention provides a method to directionally clone a DNA sequence of interest which employs a recipient vector comprising a DNA sequence of interest, e.g., optionally encoding a fusion protein, flanked by at least two restriction enzyme sites, one of which is for a first restriction enzyme that has a low frequency of recognition sites in a plurality of cDNAs or open reading frames for a particular species and generates single-strand DNA overhangs, and the other of which flanking sites is for a second restriction enzyme that has a low frequency of recognition sites in a plurality of cDNAs or open reading frames for a particular species and generates blunt ends. An acceptor vector may comprise a counter-selectable marker flanked by at least two restriction enzymes sites. One of the flanking sites in the acceptor vector is for a third restriction enzyme which generates single-strand DNA overhangs which are complementary with the single-strand DNA overhangs produced when the recipient vector is digested with the first restriction enzyme. The other flanking site in the acceptor vector is for a fourth restriction enzyme which generates blunt ends. The method includes contacting the recipient vector with the first and second restriction enzymes and the acceptor vector with the third and fourth restriction enzymes, ligating the resulting linear molecules, transforming a host cell with the ligation mixture, and selecting for host

cells with desirable recombinant molecules, i.e., vectors with the DNA sequence of interest and the acceptor vector backbone, e.g., vectors which lack the counter-selectable gene, and optionally include a selectable marker present on the acceptor vector backbone. In one embodiment, the first and third restriction enzymes are the same. In one embodiment, the second and fourth restriction enzymes are the same. In this manner, DNA sequences of interest may be moved from one expression vector to another, for instance, to express a fusion protein encoded by a fusion of acceptor vector sequences, the exchange site(s), and the DNA sequence of interest.

The methods of the invention thus employ one or more restriction enzymes that generate unique ends and optionally ligase to clone an open reading frame of interest. Vectors with one or more restriction enzyme sites for restriction enzymes that provide unique ends are particularly useful in directional cloning and ordered gene assembly. Moreover, the use of the vectors and methods of the invention is easy, inexpensive, fast, automatable, and results in high fidelity. Further, the vectors may be designed to express fusion proteins with no or one to a few, e.g., less than 7, amino acid residues fused to the N-terminus, C-terminus, or both the N- and C-termini. For instance, fusions generated with *Sfi*I sites flanking the DNA sequence of interest may yield fusion proteins with 4 amino acid residues at the N-terminus and C-terminus, while fusions generated with *Sgf*I/*Pme*I or *Sap*I sites flanking the DNA sequence of interest may yield fusion proteins with a single amino acid residue only at the C-terminus. If *Sfi*I or *Pme*I sites are added to a DNA sequence of interest, e.g., using an amplification reaction, an additional 3-5 bp flanking the recognition site may be included to increase cleavage efficiency. Moreover, N- and/or C-terminal fusions with fusion partner sequences useful in purification, e.g., immobilization, solubilization, *in situ* detection, protein domain studies, and protein-protein interactions, e.g., *in vitro* or *in vivo*, may be prepared, wherein fusion partner sequences are encoded by acceptor vector sequences and/or exchange sites.

Also provided is a recombinant host cell useful to reduce unintended expression from a vector. In one embodiment, the host cell is deficient in one or more inducible genes, for instance, the host cell does not express one or more

rhamnose catalytic genes, e.g., the host cell is *rhaBAD*⁻, and comprises an expression vector, e.g., one which is stably introduced to the host cell. The expression vector comprises an inducible promoter for the one or more genes, which promoter has a low level of uninduced expression and preferably has a relatively
5 slow induction profile but high final levels of expression, e.g., a *rhaBAD* promoter, and which promoter is operably linked to an open reading frame, such as one for a heterologous (non-native) transcription regulatory gene product, e.g., a RNA polymerase. In one embodiment, the recombinant host cell is deficient in rhamnose catabolism, and has a recombinant DNA molecule comprising a rhamnose-inducible
10 promoter operably linked to an open reading frame for a heterologous RNA polymerase. In one embodiment, the host cell is a prokaryotic cell, for instance, an *E. coli* cell. In one embodiment, the heterologous RNA polymerase is a phage RNA polymerase, such as a T7 RNA polymerase. The recombinant host cell may be contacted with an expression vector comprising a promoter for the heterologous
15 RNA polymerase and an open reading frame of interest, and rhamnose, e.g., either simultaneously or sequentially. In one embodiment, the expression vector comprises a transcription terminator sequence, e.g., *rrnB*, and a promoter 5' to the open reading frame of interest, which promoter is upregulated by the heterologous transcription regulatory gene product, as well as restriction sites for one or more
20 infrequent cutters flanking the open reading frame, and optionally, in the vector backbone, a selectable marker gene, a sequence which specifies a high vector copy number, and a sequence which reduces vector multimerization, e.g., *cer*. An expression vector comprising a promoter such as one for a heterologous transcription regulatory gene product, such as a RNA polymerase, which promoter
25 is operably linked to an open reading frame of interest, may also be employed in an *in vitro* transcription/translation system.

Further provided is an isolated nucleic acid fragment encoding barnase which lacks a secretory domain (signal), a vector comprising the nucleic acid fragment, such as one which comprises a promoter, for instance, a λP_L promoter
30 linked to the nucleic acid fragment, isolated protein encoded by the nucleic acid fragment, and a host cell comprising the vector. Optionally, the host cell expresses

barstar. In one embodiment, the host cell expresses barstar from a promoter which is constitutively expressed in prokaryotic cells. Optionally, the host cell is an *E. coli* cell. In one embodiment, an open reading frame for barstar is expressed from a 4c promoter. In one embodiment, the vector system of the invention includes a second
5 vector comprising a counterselectable gene comprising a nucleic acid fragment encoding a barnase which lacks a secretory domain.

Also provided is a vector comprising an open reading frame 3' to a DNA fragment of no more than 30 base pairs. The DNA fragment comprises a ribosome binding site, a *SgfI* recognition site, and a sequence which, when present in mRNA,
10 enhances the binding of the mRNA to the small subunit of a eukaryotic ribosome. In one embodiment, the DNA fragment includes AAGGAGCGATCGCCATGX (SEQ ID NO:1), and wherein X is A, T, G or C.

Further provided is a vector comprising a *SgfI* recognition site, a sequence which comprises ATG and which sequence, when present in mRNA, enhances the
15 binding of the mRNA to the small subunit of a eukaryotic ribosome, and an open reading frame which begins at the ATG in the sequence.

The invention includes a vector comprising a *SgfI* recognition site 5' to a recognition site for a first restriction enzyme which has infrequent restriction sites in cDNAs or open reading frames from at least one species and generates blunt ends,
20 which vector, once digested with *SgfI* and the first restriction enzyme and ligated to a DNA fragment comprising an open reading frame flanked by an end generated by a second restriction enzyme that generates a 3' TA overhang and an end generated by a third restriction enzyme that generates blunt ends, yields a recombinant vector comprising the open reading frame. In one embodiment, the first and third
25 restriction enzymes are the same. In another embodiment, the second restriction enzyme is *SgfI*.

Also provided is a vector comprising a first open reading frame which includes a *SgfI* recognition site and a recognition site which is not in the open reading frame for a restriction enzyme that has infrequent restriction sites in cDNAs
30 or open reading frames from at least one species and generates blunt ends. The vector, once digested with *SgfI* and the restriction enzyme which has infrequent

restriction sites in cDNAs or open reading frames from at least one species and generates blunt ends, and ligated to a DNA fragment comprising a second open reading flanked by a single-strand 3' TA DNA overhang and a blunt end, yields a recombinant vector comprising a third open reading frame comprising the first and
5 second open reading frames, which third open reading frame encodes a fusion peptide or protein.

The invention also includes a vector comprising a ribosome binding site which optionally overlaps by one nucleotide with a *SgfI* recognition site and a recognition site which is not in the open reading frame for a restriction enzyme that
10 has infrequent restriction sites in cDNAs or open reading frames from at least one species and generates blunt ends. The vector, once digested with *SgfI* and the restriction enzyme that has infrequent restriction sites in cDNAs or open reading frames from at least one species and generates blunt ends, and ligated to a DNA fragment comprising an open reading frame encoding a peptide or polypeptide
15 flanked by



and a blunt end, yields a recombinant vector which encodes the peptide or
20 polypeptide, wherein X_1 is the first codon which is 3' to the start codon for the open reading frame, wherein X_2 is the complement of X_1 , wherein Y_1 is the remainder of the open reading frame, and wherein Y_2 is the complement of Y_1 .

Further provided is a vector comprising a first open reading frame which includes a *PmeI* recognition site and is flanked at the 5' end by a recognition site for
25 a first restriction enzyme that generates complementary single-strand DNA overhangs, which vector, once digested with *PmeI* and the first restriction enzyme, and ligated to a DNA fragment comprising a blunt end at the 5' end of a second open reading frame and an end generated by a second restriction enzyme which generates single-strand DNA overhangs which are complementary to the single-
30 strand DNA overhangs generated by the first restriction enzyme, yields a recombinant vector comprising a third open reading frame comprising the first and second open reading frames.

The invention further includes a vector comprising a first open reading frame which includes a *PmeI* recognition site and is flanked at the 5' end by site for a first restriction enzyme that generates complementary single-strand DNA overhangs. The vector, once digested with *PmeI* and the first restriction enzyme, and ligated to a DNA fragment comprising a blunt end and an end generated by a second restriction enzyme which generates single-strand DNA overhangs which are complementary to the single-strand DNA overhangs generated by the first restriction enzyme, yields a recombinant vector which includes $N_1N_2N_3GTTTN_4N_5$, wherein $N_1N_2N_3GTTT$ is a sequence from the 3' end of the digested expression vector. In one embodiment, the triplet $N_1N_2N_3$ does not code for a stop codon, and N_4 and $N_5 = A$, or $N_4 = A$ and $N_5 = G$ or $N_4 = G$ and $N_5 = A$. In another embodiment, the triplet $N_1N_2N_3$ codes for a stop codon.

The vectors of the invention and methods of the invention which employ the vectors, are particularly useful in directional cloning of open reading frames. However, the vectors and methods of the invention are useful in other applications, for example, they may be employed to prepare probes, e.g., radioactive or nonradioactive probes, for instance, using vectors with promoters specific for a polymerase, such as bacteriophage polymerases, to prepare single-strand sense or anti-sense probes or therapeutic antisense RNA; or to insert a gene in an antisense orientation such that it is not expressed or expressed only after structural rearrangement (conditional gene inactivation), e.g., via recombination with *Cre/lox* (U.S. Patent No. 5,658,772), FLP/FRT, the Gin recombinase of Mu, the Pin recombinase of *E. coli*, and the R/RS system of the pSR1 plasmid.

Also provided is a method for performing genetic analysis. The method comprises populating a database of genetic data with genetic data to create a plurality of genetic records. The database containing genetic data is queried to identify a first subset of genetic records, wherein each record has at least one recognition site for restriction enzymes included in a set of predetermined restriction enzymes, and a set of statistics associated with the restriction enzyme recognition sites for at least a second subset of genetic records in the first subset is determined.

Further provided is a computer-readable medium having computer executable instructions for performing a method for performing genetic analysis. The medium includes populating a database of genetic data with a plurality of genetic records, querying the database of genetic data to identify a first subset of genetic records, wherein each record has at least one recognition site for one
5 predetermined restriction enzyme or for restriction enzymes included in a set of predetermined restriction enzymes, and determining a set of statistics associated with the restriction enzyme recognition sites for at least a second subset of genetic records in the first subset. Also provided is a computerized system for genetic
10 analysis. The system includes a database of genetic data, a processor, a set of one or more programs executed by the processor causing the processor to query the database of genetic data to identify a first subset of genetic records, wherein each record has at least one recognition site for one predetermined restriction enzyme or for restriction enzymes included in a set of predetermined restriction enzymes, and
15 determine a set of statistics associated with the restriction enzyme recognition sites for at least a second subset of genetic records in the first subset. In one embodiment, the set of statistics includes, e.g., includes determining, a number of genetic records including recognition sites for one predetermined restriction enzyme or for each of the predetermined restriction enzymes in the set. In one embodiment,
20 the set of statistics includes, e.g., includes determining, a number of occurrences of at least one site for the one predetermined restriction enzyme or for the predetermined restriction enzymes in a genetic record in the second subset. In one embodiment, the genetic records comprise nucleic acid sequences. In one embodiment, the method further comprises filtering, or a processor is further
25 operable to filter, the subset of genetic records to include or exclude genetic records having one or more selected characteristics. In another embodiment, the method further comprises filtering, or a processor is further operable to filter, the subset of genetic records to exclude genetic records having a size greater than a predetermined value. In one embodiment, the predetermined value is 21000
30 characters. In another embodiment, the method further comprises determining, or a processor is further operable to determine, a sequence of specific bases which are

present as ambiguous bases within a recognition site or which are present between a recognition site for a restriction enzyme and the position at which the restriction enzyme cleaves DNA containing the recognition site.

5

Brief Description of the Figures

Figure 1. Exemplary hapaxomers.

Figures 2A-B. Examples of hapaxomers with 3' or 5' overhangs. A) The symmetry of the site recognized by *AlwNI*, a restriction enzyme that cleaves an interrupted palindrome within the recognition site. If the bases denoted "N" are
10 ignored, the site is symmetrically equivalent to a *PvuII* site. Arrows indicate the cleavage sites on both strands. Note that a recognition and cleavage site on only one strand must be stipulated owing to the existence of a two-fold axis of symmetry. However, because cleavage by *AlwNI* results in DNA with overhangs consisting of three bases with four possibilities for each unspecified base, the sequence at the
15 termini will be different depending on the strand. B) The *FokI* recognition and cleavage sites illustrated in both orientations. Because the site lacks symmetry, there are two ways to write the bases from 5' to 3'. The cleavage sites on both strands, indicated by arrows, must be specified in order to indicate where cutting will occur. Because the cleavage sites are outside the recognition site, the single-stranded
20 overhangs can be any set of four bases. Note that *AlwNI* generates 3' overhangs, whereas *FokI* generates 5' overhangs.

Figure 3. A flowchart to identify restriction enzymes that have infrequent recognition sites in the genome of a particular organism.

Figure 4. Comparison of the percent of sequences in various organisms
25 which lack (0), have no or one (0-1), or no, one or two (0-2) recognition sites for *SapI*, *SfiI* or *SgfI/PmeI*.

Figure 5. Site frequencies of selected restriction enzymes in six species.

Figure 6. General overview of the use of interrupted palindromes for directional cloning.

30 Figure 7. Directional cloning using *SfiI*.

Figure 8. PCR interrupted palindromes cloning pathways.

Figures 9A-B. PCR interrupted palindromes cloning pathways.

Figures 10A-B. PCR interrupted palindromes cloning pathways.

Figure 11. Restriction endonucleases useful for directional cloning with *Sfi*I or other restriction enzymes generating 3 base 3' overhangs.

5 Figure 12. General overview of the use of Type IIS enzymes for directional cloning.

Figure 13. Directional cloning using *Sap*I.

Figures 14A-B. Two enzyme approach for directional cloning with an enzyme that generates staggered ends and an enzyme that generates blunt ends, e.g.,
10 *Sgf*I and *Pme*I.

Figure 15. Two enzyme cloning pathway with PCR entry.

Figure 16. Use of *Sgf*I to generate N-terminal fusions or no fusion at the N-terminus.

Figure 17. Use of *Pme*I to generate C-terminal fusions including fusions
15 with a single amino acid.

Figure 18. Use of a combination of *Sgf*I, *Pme*I, *Pac*I and *Swa*I to prepare a vector encoding two proteins of interest.

Figures 19A-B. N-terminal *Pac*I-*Sgf*I fusion site and C-terminal *Pme*I-*Swa*I fusion site.

20 Figure 20A. Exemplary luciferase donor and acceptor vectors of the invention.

Figure 20B. Analysis ligation of the donor and acceptor vector sequences having *Sfi*I sites flanking distinguishable luciferase genes.

Figure 21A-E. Exemplary vectors of the invention. KanR = kanamycin
25 resistance gene; AmpR= ampicillin resistance gene; ColE1 ori = origin of replication sequence; cer = XerCD site-specific recombinase target site; rrnB term = bidirectional terminator; T7 P = T7 RNA polymerase promoter; RBS/Kozak = ribosome binding site and Kozak sequences; and T7 term = T7 RNA polymerase termination sequence.

30 Figure 22A. Luciferase expression after induction of expression in 3 different hosts at 37°C.

Figure 22B. Luciferase expression in 3 different hosts at 25°C, t = 0.

Figure 22C. Luciferase expression in 3 different hosts at 25°C, t = 5 hours and 21 hours.

5

Detailed Description of the Invention

Definitions

The term "unique restriction enzyme site" indicates that the recognition sequence for a given restriction enzyme appears once within a nucleic acid molecule.

10

The terms "polylinker" or "multiple cloning site" refer to a cluster of restriction enzyme sites on a nucleic acid construct which are utilized for the insertion and/or excision of nucleic acid sequences such as the coding region of a gene, *lox* sites, etc.

15

The term "prokaryotic termination sequence" refers to a nucleic acid sequence which is recognized by the RNA polymerase of a prokaryotic host cell and results in the termination of transcription. Prokaryotic termination sequences commonly comprise a GC-rich region that has a twofold symmetry followed by an AT-rich sequence. Commonly used prokaryotic termination sequences are the T7 and *rrnB* termination sequences. A variety of termination sequences are known to the art and may be employed in the nucleic acid constructs of the present invention including, the T_{INT}, T_{L1}, T_{L2}, T_{L3}, T_{R1}, T_{R2}, T_{6S} termination signals derived from the bacteriophage lambda and termination signals derived from bacterial genes such as the *trp* gene of *E. coli*.

20

The term "eukaryotic polyadenylation sequence" (also referred to as a "poly A site" or "poly A sequence") as used herein denotes a DNA sequence which directs both the termination and polyadenylation of the nascent RNA transcript. Efficient polyadenylation of the recombinant transcript is desirable as transcripts lacking a poly A tail are unstable and are rapidly degraded. The poly A signal utilized in an expression vector may be "heterologous" or "endogenous." An endogenous poly A signal is one that is found naturally at the 3' end of the coding region of a given gene in the genome. A heterologous poly A signal is one which is isolated

25

30

from one gene and placed 3' of another gene. A commonly used heterologous poly A signal is the SV40 poly A signal. The SV40 poly A signal is contained on a 237 bp *Bam*HI/*Bcl*II restriction fragment and directs both termination and polyadenylation (Sambrook et al., Molecular Cloning: A Laboratory Manual, Cold Spring Harbor (1989)); numerous vectors contain the SV40 poly A signal. Another commonly used heterologous poly A signal is derived from the bovine growth hormone (BGH) gene; the BGH poly A signal is available on a number of commercially available vectors. The poly A signal from the herpes simplex virus thymidine kinase (HSV tk) gene is also used as a poly A signal on expression vectors.

As used herein, the terms "selectable marker" or "selectable marker gene" refers to the use of a gene which encodes an enzymatic activity that confers the ability to grow in medium lacking what would otherwise be an essential nutrient (e.g., the TRP1 gene in yeast cells); in addition, a selectable marker may confer resistance to an antibiotic or drug upon the cell in which the selectable marker is expressed. A selectable marker may be used to confer a particular phenotype upon a host cell. When a host cell must express a selectable marker to grow in selective medium, the marker is said to be a positive selectable marker (e.g., antibiotic resistance genes which confer the ability to grow in the presence of the appropriate antibiotic). Selectable markers can also be used to select against host cells containing a particular gene (e.g., the *sacB* gene which, if expressed, kills the bacterial host cells grown in medium containing 5% sucrose); selectable markers used in this manner are referred to as negative selectable markers or counter-selectable markers.

As used herein, the term "vector" is used in reference to nucleic acid molecules that transfer DNA segment(s) from one cell to another. The term "vehicle" is sometimes used interchangeably with "vector." A "vector" is a type of "nucleic acid construct." The term "nucleic acid construct" includes circular nucleic acid constructs such as plasmid constructs, plasmid constructs, cosmid vectors, etc. as well as linear nucleic acid constructs (e.g., lambda, phage constructs, PCR products). The nucleic acid construct may comprise expression signals such as a

promoter and/or an enhancer (in such a case it is referred to as an expression vector).

5 The term "expression vector" as used herein refers to a recombinant DNA molecule containing a desired coding sequence and appropriate nucleic acid sequences necessary for the expression of the operably linked coding sequence in a particular host organism. Nucleic acid sequences necessary for expression in procaryotes usually include a promoter, an operator (optional), and a ribosome binding site, often along with other sequences. Eukaryotic cells are known to utilize promoters, enhancers, and termination and polyadenylation signals.

10 The terms "in operable combination", "in operable order" and "operably linked" as used herein refer to the linkage of nucleic acid sequences in such a manner that a nucleic acid molecule capable of directing the transcription of a given gene and/or the synthesis of a desired protein molecule is produced. The term also refers to the linkage of amino acid sequences in such a manner so that a functional
15 protein is produced.

The terms "transformation" and "transfection" as used herein refer to the introduction of foreign DNA into prokaryotic or eucaryotic cells. Transformation of prokaryotic cells may be accomplished by a variety of means known to the art including the treatment of host cells with CaCl_2 to make competent cells,
20 electroporation, etc. Transfection of eukaryotic cells may be accomplished by a variety of means known to the art including calcium phosphate-DNA co-precipitation, DEAE-dextran-mediated transfection, polybrene-mediated transfection, electroporation, microinjection, liposome fusion, lipofection, protoplast fusion, retroviral infection, and biolistics.

25 As used herein, the terms "restriction endonucleases" and "restriction enzymes" refer to bacterial enzymes, each of which cut double-stranded DNA at or near a specific nucleotide sequence.

As used herein, the term "recombinant DNA molecule" as used herein refers to a DNA molecule which is comprised of segments of DNA joined together by
30 means of molecular biological techniques.

As used herein, "recognition site" refers to a sequence of specific bases that is recognized by a restriction enzyme if the sequence is present in double-stranded DNA; or, if the sequence is present in single-stranded RNA, the sequence of specific bases that would be recognized by a restriction enzyme if the RNA was reverse transcribed into cDNA and the cDNA employed as a template with a DNA polymerase to generate a double-stranded DNA; or, if the sequence is present in single-stranded DNA, the sequence of specific bases that would be recognized by a restriction enzyme if the single-stranded DNA was employed as a template with a DNA polymerase to generate a double-stranded DNA; or, if the sequence is present in double-stranded RNA, the sequence of specific bases that would be recognized by a restriction enzyme if either strand of RNA was reverse transcribed into cDNA and the cDNA employed as a template with a DNA polymerase to generate a double-stranded DNA.

An "open reading frame" includes at least 3 consecutive codons which are not stop codons.

DNA molecules are said to have "5' ends" and "3' ends" because mononucleotides are reacted to make oligonucleotides in a manner such that the 5' phosphate of one mononucleotide pentose ring is attached to the 3' oxygen of its neighbor in one direction via a phosphodiester linkage. Therefore, an end of an oligonucleotides referred to as the "5' end" if its 5' phosphate is not linked to the 3' oxygen of a mononucleotide pentose ring and as the "3' end" if its 3' oxygen is not linked to a 5' phosphate of a subsequent mononucleotide pentose ring. As used herein, a nucleic acid sequence, even if internal to a larger oligonucleotide, also may be said to have 5' and 3' ends. In either a linear or circular DNA molecule, discrete elements are referred to as being "upstream" or 5' of the "downstream" or 3' elements. This terminology reflects the fact that transcription proceeds in a 5' to 3' fashion along the DNA strand. The promoter and enhancer elements which direct transcription of a linked gene are generally located 5' or upstream of the coding region. However, enhancer elements can exert their effect even when located 3' of the promoter element and the coding region. Transcription termination and polyadenylation signals are located 3' or downstream of the coding region.

As used herein, the term "an oligonucleotide having a nucleotide sequence encoding a gene" means a nucleic acid sequence comprising the coding region of a gene or in other words the nucleic acid sequence which encodes a gene product. The coding region may be present in either a cDNA, genomic DNA or RNA form. When
5 present in a DNA form, the oligonucleotide may be single-stranded (i.e., the sense strand) or double-stranded. Suitable control elements such as enhancers/promoters, splice junctions, polyadenylation signals, etc. may be placed in close proximity to the coding region of the gene if needed to permit proper initiation of transcription and/or correct processing of the primary RNA transcript. Alternatively, the coding
10 region utilized in the vectors of the present invention may contain endogenous enhancers/promoters, splice junctions, intervening sequences, polyadenylation signals, etc. or a combination of both endogenous and exogenous control elements.

As used herein, the term "regulatory element" refers to a genetic element which controls some aspect of the expression of nucleic acid sequences. For
15 example, a promoter is a regulatory element which facilitates the initiation of transcription of an operably linked coding region. Other regulatory elements include splicing signals, polyadenylation signals, termination signals and the like.

Transcriptional control signals in eukaryotes comprise "promoter" and "enhancer" elements. Promoters and enhancers consist of short arrays of DNA
20 sequences that interact specifically with cellular proteins involved in transcription (Maniatis et al., Science, 236:1237 (1987)). Promoter and enhancer elements have been isolated from a variety of eukaryotic sources including genes in yeast, insect and mammalian cells and viruses (analogous control elements, i.e., promoters, are also found in prokaryotes). The selection of a particular promoter and enhancer
25 depends on what cell type is to be used to express the protein of interest. Some eukaryotic promoters and enhancers have a broad host range while others are functional in a limited subset of cell types (for review see Voss et al., Trends Biochem. Sci., 11:287 (1986) and Maniatis et al., *supra* (1987)). For example, the SV40 early gene enhancer is very active in a wide variety of cell types from many
30 mammalian species and has been widely used for the expression of proteins in mammalian cells (Dijkema et al., EMBO J., 4:761 (1985)). Two other examples of

promoter/enhancer elements active in a broad range of mammalian cell types are those from the human elongation factor 10 gene (Uetsuki et al., J. Biol. Chem., 264:5791 (1989), Kim et al., Gene, 91:217 (1990) and Mizushima et al., Nuc. Acids. Res., 18:5322 (1990)) and the long terminal repeats of the Rous sarcoma virus (Gorman et al., Proc. Natl. Acad. Sci. USA, 79:6777 (1982)) and the human cytomegalovirus (Boshart et al., Cell, 41:521 (1985)).

As used herein, the term "promoter/enhancer" denotes a segment of DNA which contains sequences capable of providing both promoter and enhancer functions (i.e., the functions provided by a promoter element and an enhancer element, see above for a discussion of these functions). For example, the long terminal repeats of retroviruses contain both promoter and enhancer functions. The enhancer/promoter may be "endogenous" or "exogenous" or "heterologous." An "endogenous" enhancer/promoter is one which is naturally linked with a given gene in the genome. An "exogenous" or "heterologous" enhancer/promoter is one which is placed in juxtaposition to a gene by means of genetic manipulation (i.e., molecular biological techniques) such that transcription of that gene is directed by the linked enhancer/promoter.

The presence of "splicing signals" on an expression vector often results in higher levels of expression of the recombinant transcript. Splicing signals mediate the removal of introns from the primary RNA transcript and consist of a splice donor and acceptor site (Sambrook et al., Molecular Cloning: A Laboratory Manual, 2nd ed., Cold Spring Harbor Laboratory Press, New York (1989) pp. 16.7-16.8). A commonly used splice donor and acceptor site is the splice junction from the 16S RNA of SV40.

Efficient expression of recombinant DNA sequences in eucaryotic cells requires expression of signals directing the efficient termination and polyadenylation of the resulting transcript. Transcription termination signals are generally found downstream of the polyadenylation signal and are a few hundred nucleotides in length. The term "poly A site" or "poly A sequence" as used herein denotes a DNA sequence which directs both the termination and polyadenylation of the nascent RNA transcript. Efficient polyadenylation of the recombinant transcript

is desirable as transcripts lacking a poly A tail are unstable and are rapidly degraded. The poly A signal utilized in an expression vector may be "heterologous" or "endogenous." An endogenous poly A signal is one that is found naturally at the 3' end of the coding region of a given gene in the genome. A heterologous poly A
5 signal is one which is one which is isolated from one gene and placed 3' of another gene.

Eukaryotic expression vectors may also contain "viral replicons" or "viral origins of replication." Viral replicons are viral DNA sequences which allow for the extrachromosomal replication of a vector in a host cell expressing the appropriate
10 replication factors. Vectors which contain either the SV40 or polyoma virus origin of replication replicate to high copy number (up to 10^4 copies/cell) in cells that express the appropriate viral T antigen. Vectors which contain the replicons from bovine papillomavirus or Epstein-Barr virus replicate extrachromosomally at low copy number (about 100 copies/cell).

15 As used herein, the terms "nucleic acid molecule encoding," "DNA sequence encoding," and "DNA encoding" refer to the order or sequence of deoxyribonucleotides along a strand of deoxyribonucleic acid. The order of these deoxyribonucleotides determines the order of amino acids along the polypeptide (protein) chain. The DNA sequence thus codes for the amino acid sequence.

20 As used herein, the term "gene" means the deoxyribonucleotide sequences comprising the coding region of a gene, e.g., a structural gene, and the including sequences located adjacent to the coding region on both the 5' and 3' ends for a distance of about 1 kb on either end such that the gene corresponds to the length of the full-length mRNA. The sequences which are located 5' of the coding region and
25 which are present on the mRNA are referred to as 5' non-translated sequences. The sequences which are located 3' or downstream of the coding region and which are present on the mRNA are referred to as 3' non-translated sequences; these sequences. The term "gene" encompasses both cDNA and genomic forms of a gene. A genomic form or clone of a gene contains the coding region interrupted with non-
30 coding sequences termed "introns" or "intervening regions" or "intervening sequences." Introns are segments of a gene which are transcribed into nuclear RNA

(hnRNA); introns may contain regulatory elements such as enhancers. Introns are removed or "spliced out" from the nuclear or primary transcript; introns therefore are absent in the messenger RNA (mRNA) transcript. The mRNA functions during translation to specify the sequence or order of amino acids in a nascent polypeptide.

5 In addition to containing introns, genomic forms of a gene may also include sequences located on both the 5' and 3' end of the sequences which are present on the RNA transcript. These sequences are referred to as "flanking" sequences or regions (these flanking sequences are located 5' or 3' to the non-translated sequences present on the mRNA transcript). The 5' flanking region may contain regulatory
10 sequences such as promoters and enhancers which control or influence the transcription of the gene. The 3' flanking region may contain sequences which direct the termination of transcription, post-transcriptional cleavage and polyadenylation.

As used herein, the term "purified" or "to purify" refers to the removal of contaminants from a sample.

15 The term "recombinant DNA molecule" as used herein refers to a DNA molecule which is comprised of segments of DNA joined together by means of molecular biological techniques.

The term "recombinant protein" or "recombinant polypeptide" as used herein refers to a protein molecule which is expressed from a recombinant DNA molecule.

20 The term "native protein" as used herein to indicate that a protein does not contain amino acid residues encoded by vector sequences; that is the native protein contains only those amino acids found in the protein as it occurs in nature. A native protein may be produced by recombinant means or may be isolated from a naturally occurring source.

25 As used herein the term "portion" when in reference to a protein (as in "a portion of a given protein") refers to fragments of that protein. The fragments may range in size from two or more amino acid residues to the entire amino acid sequence minus one amino acid.

As used herein, the term "fusion protein" refers to a chimeric protein
30 containing the protein of interest joined to a different peptide or protein fragment. The fusion partner may, for example, enhance the solubility of a linked protein of

interest, may provide an epitope tag or affinity domain to allow identification and/or purification of the recombinant fusion protein, e.g., from a host cell which expresses the fusion or a culture supernatant of that cell, or both, or may have another property or activity, e.g., two functional enzymes can be fused to produce a single protein with multiple enzymatic activities. If desired, the fusion protein may be removed from the protein of interest by a variety of enzymatic or chemical means known to the art. Thus, examples of fusion protein producing sequences useful in the vectors of the invention include epitope tag encoding sequences, affinity domain encoding sequences, or other functional protein encoding sequences, and the like. The use of the term "functional protein encoding sequence", as used herein, indicates that the fusion protein producing element of a vector encodes a protein or peptide having a particular activity, such as an enzymatic activity, e.g., luciferase or dehalogenase, a binding activity, and the like, e.g., thioredoxin. For example, a functional protein encoding sequence may encode a kinase catalytic domain (Hanks and Hunter, FASEB J, 9:576-595, 1995), producing a fusion protein that can enzymatically add phosphate moieties to particular amino acids, or may encode a Src Homology 2 (SH2) domain (Sadowski, et al., Mol. Cell. Bio., 6:4396, 1986; Mayer and Baltimore, Trends Cell. Biol., 3:8, 1993), producing a fusion protein that specifically binds to phosphorylated tyrosines.

20 I. Restriction Enzyme Sites and Enzymes Useful in the Vectors and Methods of the Invention

The present invention employs two general approaches to directional cloning and ordered gene assembly. In one approach, restriction sites for hapaxotermistic restriction enzymes, e.g., those with degenerate recognition or cleavage sequences (see Figures 1-2), are employed. Hapaxotermistic enzymes are enzymes able to generate unique ends (Table 1). *FokI*, a type IIS enzyme, is included and so is *AlwNI*, an interrupted palindrome. Because the cleavage site is located among the unspecified bases, the termini are expressed in N's. Unless the complete nucleotide sequence within the interruption or flanking the recognition site is written, the detailed nature of the ends cannot be stated; statistically speaking, all single stranded overhangs will be different. It is also unlikely that these overhangs possess

elements of symmetry. In the general case, this means that the protruding bases are not composed of an asymmetric unit followed by its reverse complement; the ends will not be self-complementary; and it will not be possible to form concatamers with a fragment bearing such ends. With nonhapaxotermistic enzymes such as *EcoRI*
5 the opposite situation prevails; both the recognition site, G↓AATTC, and the overhanging ends produced by cleavage, AATT, always display palindrome-like elements; and the overhang of any fragment is complementary with itself and with the protruding ends of all other fragments generated by the same enzyme.

Table 1

<i>Alw</i> NI	↓ CAGNNNCTG GTCNNNGAC ↑	<i>Dra</i> III	↓ CACNNNGTG GTGNNNCAC ↑
<i>Bbs</i> I	↓ GAAGACNN CTTCTGNNNNNN ↑	<i>Ear</i> I	↓ CTCTTCN GAGAAGNNNN ↑
<i>Bbv</i> I	↓ GCAGCNNNNNNNN CGTCGNNNNNNNNNNNN ↑	<i>Esp</i> 3 I	↓ CGTCTCN GCAGAGNNNNNN ↑
<i>Bgl</i> I	↓ GCCNNNNNGGC CGGNNNNNCCG ↑	<i>Fok</i> I	↓ GGATGNNNNNNNNNN CCTACNNNNNNNNNNNN ↑
<i>Bsa</i> I	↓ GGTCTCN CCAGAGNNNNNN ↑	<i>Hga</i> I	GACGCNNNNNN CTGCGNNNNNNNNNNNN
<i>Bsl</i> I	↓ CCNNNNNNNGG GGNNNNNNNCC ↑	<i>Mwo</i> I	↓ GCNNNNNNNGC CGNNNNNNNCG ↑
<i>Bsm</i> AI	↓ GTCTCN CAGAGNNNNNN ↑	<i>Pfl</i> MI	↓ CCANNNNNNTGG GGTNNNNNACC ↑
<i>Bsm</i> FI	↓ GTCCNNNNNNNNNN CAGGGNNNNNNNNNNNN ↑	<i>Sap</i> I	↓ GCTCTTCN CGAGAAGNNNN ↑
<i>Bsp</i> MI	↓ ACCTGNNNN TGGACGNNNNNNNN ↑	<i>Sfa</i> NI	↓ GCATCNNNN CGTAGNNNNNNNNNN ↑
<i>Bst</i> XI	↓ CCANNNNNNTGG GGTNNNNNACC ↑	<i>Sfi</i> I	↓ GGCCNNNNNGGCC CCGGNNNNNCCGG ↑

Note. The cleavage sites are indicated by the arrows. Isoschizomers occur in several cases. The enzymes listed and their isoschizomers are as follows: *Bbs*I, *Bsc*91I; *Bbv*I, *Bst*I, *Bst*71I; *Bsa*I, *Eco*31I, *Bsm*AI, *Alw*26I; *Ear*I, *Ksp*632I; and *Pfl*MI, *Acc*B7I.

Enzymes which generate blunt ends can never be hapaxotermistic. For instance, the restriction site for *Bsa*BI has N's but the enzyme produces blunt end.

There are enzymes that are formally, but not functionally, hapaxomers. In this category are restriction endonucleases that generate overhangs of only one or two unspecified bases such as *Alw*I and *Bpm*I, respectively (Table 2). Conversely, those type II enzymes which recognize sites with multiple degeneracies are functionally, but not formally, hapaxomers. For example, if a fragment were to be cut at several locations by *Bsp*1286I (Table 2), an array of single stranded extensions, e.g., GGCC, TGCA, AGCT, GGCA, GGCT, AGCC, AGCA, TGCC, and TGCT, might occur. The first three of these possess an obvious element of symmetry which eliminates them from consideration. The last six protrusions do not possess an element of symmetry and, therefore, are neither self-complementary nor self-ligatable; they have the potential to be unique. On that basis *Bsp*1286I is a hapaxomer. Hapaxotermicity is the ability to generate a finite percentage of overhangs lacking in symmetry. The symmetry or lack thereof of the restriction enzyme recognition site is of no consequence.

Table 2

<i>Alw</i> I	↓ GGATCNNNN CCTAGNNNNN ↑	A hapaxomer with an overhang of one base
<i>Bpm</i> I	↓ CTGGAGNNNNNNNNNNNNNNNNNNNN GACCTCNNNNNNNNNNNNNNNNNNNN ↑	A hapaxomer with an overhang of two bases
<i>Bsp</i> 1286 I	G C ↓ G A G C A C T T C G C T C G T G ↑A A	An honorary hapaxomer

*Bsp*1286I has overhangs of four bases on each strand; two bases are uniquely specified and two are restricted to one of three possibilities. Clearly, the statistical

probability that the ends are unique is less than that of enzymes which generate two completely unspecified overhanging bases. Such enzymes include *BcgI*, *BpmI*, *BsaII*, *BsgI*, *BsrDI*, *DrdI*, and *Eco57I*.

In one embodiment of the invention, a donor vector is obtained or prepared.

- 5 The donor vector includes a DNA sequence of interest flanked by at least two restriction enzyme sites, at least one of which is for a first restriction enzyme with a degenerate recognition sequence. In another embodiment, the DNA sequence of interest is flanked by two restriction enzyme sites for a restriction enzyme with a degenerate recognition sequence, which sites are not identical and so, once the
- 10 donor vector is cleaved with that enzyme, yields a linear DNA with non-self complementary single-strand DNA overhangs. The donor vector also contains at least one selectable marker gene which optionally is not the DNA sequence of interest, e.g., the selectable marker gene is part of the vector backbone. The donor vector is useful to transfer the DNA sequence of interest in an oriented manner to an
- 15 acceptor vector for expression of the DNA sequence of interest in the resulting recipient vector. The acceptor vector contains non-essential DNA sequences flanked by at least two restriction enzyme sites for a second restriction enzyme with a degenerate recognition sequence which yields non-self complementary single-strand DNA overhangs. Those sites, once cleaved, yield single-strand DNA
- 20 overhangs that are each complementary to only one of the two single-strand DNA overhangs generated by the first restriction enzyme. In one embodiment, the first and second restriction enzymes are the same. In another embodiment, the first and second restriction enzymes are different and are not isoschizomers and so, the resulting ligated sequences (the exchange site) are not cleavable by at least one of
- 25 restriction enzymes having a degenerate recognition sequence that is employed to transfer the DNA sequence of interest. For example, the fusion of single-strand DNA overhangs generated by *BglI* and single-strand DNA overhangs generated by *SfiI* results in an exchange site that is not cleavable by *SfiI*, but is cleavable by *BglI*. Similarly, the fusion of single-strand DNA overhangs generated by *SgfI* and single-
- 30 strand DNA overhangs generated by *PvuI* results in an exchange site that is not cleavable by *SgfI*, but is cleavable by *PvuI*. Further, the fusion of ends generated by

PmeI and ends generated by *DraI* results in an exchange site that is not cleavable by *PmeI*, but is cleavable by *DraI*.

In another approach, a donor vector is obtained or prepared that contains a DNA sequence of interest flanked by at least two restriction enzyme sites, one of which is for a first restriction enzyme which has infrequent restriction sites in cDNAs or open reading frames from at least one species and generates single-strand DNA overhangs, and another of which is for a second restriction enzyme that has infrequent restriction sites in cDNAs or open reading frames from at least one species and generates ends that are not complementary to the overhangs generated by the first restriction enzyme. In one embodiment, the second restriction enzyme generates blunt ends. The donor vector also contains at least one selectable marker gene which optionally is not the DNA sequence of interest. The donor vector is useful to transfer the DNA sequence of interest in an oriented manner to an acceptor vector for expression of the DNA sequence of interest, resulting in a recipient vector. The acceptor vector contains non-essential DNA sequences flanked by at least two restriction enzyme sites. In one embodiment, the non-essential DNA sequences comprise a counter-selectable gene, e.g., a barnase gene, a *ccdB* gene, or a *SacB* gene. One of the flanking restriction sites in the acceptor vector is for a third restriction enzyme which generates single-strand DNA overhangs, which overhangs are complementary to the single-strand DNA overhangs produced by digestion of the donor vector with the first restriction enzyme. In one embodiment, the restriction site for the third restriction enzyme is different than the restriction site for the first restriction enzyme and the sites are not cleaved by the same restriction enzyme. In another embodiment, the first and third restriction enzymes are the same. The other flanking restriction site in the acceptor vector is for a fourth restriction enzyme which yields ends that are not complementary to the ends generated by the first or third restriction enzyme. In one embodiment, the second and fourth restriction enzymes generate blunt ends. In one embodiment, the restriction site for the fourth restriction enzyme is different than the restriction site for the second restriction enzyme and the sites are not cleaved by the same restriction enzyme. In this manner, the exchange site is likely not cleavable by the

second or fourth restriction enzyme. In another embodiment, the second and fourth restriction enzymes are the same.

Thus, by designing a donor vector and an acceptor vector with selected restriction enzyme sites which are appropriately positioned, once these vectors are digested with the respective restriction enzymes, the DNA sequence of interest can only be oriented in one direction in the acceptor vector backbone.

Restriction enzyme sites useful in the practice of the invention include but are not limited to hapaxomeric sequences, sequences recognized by class II enzymes or class IIS enzymes, as well as restriction enzyme sites recognized by enzymes that yield blunt ends, and including enzymes that are infrequent cutters in one or more species.

Suitable class IIS restriction enzymes include those enzymes that recognize a five-base contiguous sequence, including but not limited to the following enzymes and their isoschizomers, which are indicated in parentheses: *Alw26I* (*BsmAI*), *AlwI* (*AcIWI*, *BinI*), *AsuHPI* (*HphI*), *BbvI* (*Bst71I*), *BceFI*, *BstF5I* (*BseGI*, *FokI*), *FauI*, *HgaI*, *MboII*, *PleI*, *SfaNI*, and *TspRI*; that recognize a six-base contiguous sequence including but not limited to the following enzymes and their isoschizomers: *AceIII*, *BbsI* (*BbvII*, *BpiI*, *BpuAI*), *Bce83I*, *BciVI*, *BfiI* (*BmrI*), *BpmI* (*GsuI*), *BsaI* (*Eco31I*), *BseRI*, *BsgI*, *BsmBI* (*Esp3I*), *BsmFI*, *BspMI*, *BsrDI* (*Bse3DI*), *Bsu6I* (*Eam1104I*, *EarI*, *Ksp632I*), *Eco57I*, *FauI*, *MmeI*, *RleAI*, *TaqII*, and *Tth111 II*. *SapI*, and its isoschizomer *VapK32I*, which recognize a seven-base sequence, and *SfiI*, which recognizes an eight-base sequence, also can be used. Further examples of useful enzymes include those that recognize a four-base pair split sequence (e.g., *Bse4I* (*BseLI*, *MsiYI*, *BsII*), *MwoI*), and enzymes that recognize a six-base pair split sequence (e.g., *AccB7I* (*Esp1396I*, *PflMI*, *Van91I*), *AdeI* (*DraIII*), *AhdI* (*AspEI*, *Eam1105I*, *EchHKL*, *NruGI*), *AlwNI*, *ApaBI* (*BstAPI*), *AspI* (*PflFI*, *Tth111I*), *BglI*, *BstXI*, *DrdI* (*DseDI*), and *EcoNI* (*XagI*, *XcmI*). Additional suitable class IIS restriction enzymes are known to those of skill in the art (see, for example, Szybalski et al., Gene, 100:13 (1991)).

There are other enzymes that are not class IIS enzymes, which produce non-palindromic ends. Examples of such enzymes include but are not limited to *AvaI*

(*Ama87I*, *BcoI*, *BsoBI*, *Eco88I*), *AvaII* (*Eco47I*, *Bme18I*, *HgiEI*, *SinI*), *BanI* (*AccB1I*, *BshNI*, *Eco64I*), *BfmI* (*BstSFI*, *SfcI*), *Bpu10I*, *BsaMI* (*BscCI*, *BsmI*, *Mva1269I*), *Bsh1285I* (*BsaOI*, *BsiEI*, *BstMCI*), *Bse1I* (*BseNI*, *BsrI*, *Cfr10I*), *BsiI* (*BssSI*, *Bst2BI*), *BsiZI* (*AspS9I*, *Cfr13I*, *Sau96I*), *Bsp1720I* (*BlpI*, *Bpu1102I*, *CellI*),
5 *Bst4CI*, *BstDEI* (*DdeI*), *CpoI* (*CspI*, *RsrII*), *DsaI* (*BstDSI*), *Eco24I* (*BanII*, *EcoT38I*, *FriOI*, *HgiIII*), *Eco130I* (*StyI*, *BssT1I*, *EcoT14I*, *ErhI*), *EspI* (*BlpI*, *Bpu1102I*, *Bsp1720I*, *CellI*), *HgiAI* (*BsiHKAI*, *Alw21I*, *AspHI*, *Bbv12I*), *HinfI*, *PspPPI* (*PpuMI*, *Psp5II*), *SanDI*, *SduI* (*Bsp1286I*, *BmyI*), *SecI* (*BsaJI*, *BseDI*), *SfcI* (*BfmI*, *BstSFI*), and *SmlI*.

10 Other enzymes useful in the invention are those which have few recognition sites in DNA, e.g., cDNAs, of one or more organisms (an “infrequent cutter”). To select restriction enzyme sites for this embodiment of the invention, analyses of sequences for a plurality of mRNAs, open reading frames and/or cDNAs from an organism are conducted, e.g., using computer software, to determine the relative
15 frequency of those sites in that organism (see Figures 3-5). For example, *SapI* has numerous recognition sites in human cDNAs, e.g., 38-43%, while the combination of *SgfI* and *PmeI*, and *SfiI*, have relatively few recognition sites in human cDNAs, for instance, 2 to 3%, and 13 to 14%, respectively. Enzymes which may generate ends complementary to *SgfI* include but are not limited to *Bce83I* (*BpuEI*), *BseMII*,
20 *BseRI*, *BsgI*, *BspCNI*, *BsrDI* (*Bse3DI*, *BseMI*), *BstF5I* (*BseGI*), *BtsI*, *DrdI* (*AasI*, *DseDI*), *EciI*, *Eco57I* (*AcuI*, *BspKT5I*), *Eco57MI*, *GsuI* (*BpmI*), *MmeI*, *TaqII*, *TspDTI*, *TspGWI*, *Tth111II*, *BspKT6I* (*BstKTI*), *PacI*, *PvuI* (*Afa22MI*, *Afa16RI*, *BspCI*, *EagBI*, *ErhB9I*, *MvrI*, *NbII*, *Ple19I*, *Psu161I*, *RshI*, *XorII*), and *SgfI* (*AsiSI*).

Enzymes which generate blunt ends include but are not limited to *AhaIII*
25 (*DraI*, *PauAII*, *SruI*), *AluI* (*MltI*), *BalI* (*MlsI*, *Mlu31I*, *MluNI*, *MscI*, *Msp20I*), *BfrBI*, *BsaAI* (*BstBAI*, *MspYI*, *PsuAI*), *BsaBI* (*Bse8I*, *BseJI*, *Bsh1365I*, *BsiBI*, *BsrBRI*, *MamI*), *BsrBI* (*AccBSI*, *BstD102I*, *Bst31NI*, *MbiI*), *BtrI* (*BmgBI*), *Cac8I* (*BstC8I*), *CdiI*, *CviJI* (*CviTI*), *CviRI* (*HpyCH4V*, *HpyF44III*), *Eco47III* (*AfeI*, *AitI*, *Aor51HI*, *FunI*), *Eco78I* (*EgeI*, *EheI*, *SfoI*), *EcoICRI* (*BpuAmI*, *Ecl136II*, *Eco53kI*, *MxaI*),
30 *EcoRV* (*CeqI*, *Eco32I*, *HjaI*, *HpyCI*, *NsiCI*), *EsaBC3I*, *FnuDII* (*AccII*, *BceBI*, *BepI*, *Bpu95I*, *Bsh1236I*, *Bsp50I*, *Bsp123I*, *BstFNI*, *BstUI*, *Bsu1532I*, *BtkI*, *Csp68KVI*,

CspKVI, FalII, FauBII, MvnI, ThalI, FspAI, HaeI, HaeIII (BanAI, BecAII, Bim19II, Bme361I, BseQI, BshI, BshFI, Bsp211I, BspBRI, BspKI, BspRI, BsuRI, BteI, CtlI, DsaII, EsaBC4I, FnuDI, MchAII, MfoAI, NgoPII, NspLKI, PalI, Pde133I, PflKI, PhoI, PlaI, SbvI, SfaI, SuaI), HindII (HinJCI, HincII), HpaI
5 *(BstEZ359I, BstHPI, KspAI, SsrI), Hpy8I (HpyBII), LpnI (Bme142I), MlyI (SchI), MslI (SmiMI), MstI (Acc16I, AoiI, AviII, FdiII, FspI, NsbI, PamI, Pun14627I), NaeI (CcoI, PdiI, SauBMKI, SauHPI, SauLPI, SauNI, SauSI, Slu1777I, SspCI), NlaIV (AspNI, BscBI, BspLI, PspN4I), NruI (Bsp68I, MluB2I, Sbo13I, SpoI), NspBII (MspA1I), OliI (AleI), PmaCI (AcvI, BbrPI, BcoAI, Eco72I, PmlI), PmeI (MssI),*
10 *PshAI (BoxI, BstPAI), PsiI, PvuII (BavI, BavAI, BavBI, Bsp153AI, BspM39I, BspO4I, Cfr6I, DmaI, EcII, NmeRI, Pae17kI, Pun14627II, Pvu84II, Uba153AI, UbaM39I), RsaI (AfaI, HpyBI, PlaAII), ScaI (Acc113I, AssI, DpaI, Eco255I, RflFII), SciI, SmaI (CfrJ4I, PaeBI, PspALI), SnaBI (BstSNI, Eco105I), SrfI, SspI, SspD5I, StuI (AatI, AspMI, Eco147I, GdiI, PceI, Pme55I, SarI, Sru30DI, SseBI,*
15 *SteI), SwaI (BstRZ246I, BstSWI, MspSWI, SmiI), XcaI (BspM90I, BssNAI, Bst1107I, BstBSI, BstZ17I), XmnI (Asp700I, BbvAI, MroXI, PdmI), and ZraI.*

In one embodiment, the restriction enzyme site in a vector of the invention is for a restriction enzyme that generates blunt ends and preferably has relatively few recognition sites in a particular organism, e.g., *PmeI (MssI), NruI (Bsp68I, MluB2I,*
20 *Sbo13I, SpoI), SnaBI (BstSNI, Eco105I), SrfI, and SwaI (BstRZ246I, BstSWI, MspSWI, SmiI),* as well as *HpaI, HincII, PshAI, Oli I, AluI, Alw26I, BalI, DraI, DpnI, EcoR47III, EcoRCRI, EcoRV, FokI, HaeIII, HincII, MboI, MspA1I, NaeI, RsaI, PvuII, ScaI, SmaI, SspI, StuI, XmnI, EcaBC3I, SciI, HincII, DraI, BsaBI, Cac8I, Hpy8I, MlyI, PshAI, SspD5I, BfrBI, BsaAI, BsrBI, BtrI, CdiI, CviJI, CviRI,*
25 *Eco47III, Eco78I, EcoICRI, FnuDII, FspAI, HaeI, LpnI, MlyI, MslI, MstI, NaeI, NlaIV, NruI, NspBII, OliI, PmaCI, PshAI, PsiI, SrfI, StuI, XcaI, XmnI, ZraI* or an isoschizomer thereof.

II. Methods to Identify Frequencies of Recognition Sites

Figure 3 is a flowchart of a method 300 for performing a genetic analysis
30 according to an embodiment of the invention. The method may be performed by one or more computer programs or modules made up of computer-executable

instructions. Describing the method by reference to a flowchart enables one skilled in the art to develop such programs or modules including such instructions to carry out the method on suitable computers (the processor or processors of the computer executing the instructions from computer-readable media such as RAM, ROM, CD-ROM, DVD-ROM, hard-drives, floppy drives and other such media). The method illustrated in Figure 3 is inclusive of acts that may be taken by an operating environment executing an exemplary embodiment of the invention.

A system executing the method begins by populating a database with genetic records obtained from a source database (block 302). Populating a database may be performed using some manual manipulations. In some embodiments, the genetic records comprise gene sequences having open reading frames, e.g., from cDNAs, or a portion thereof. In some embodiments, the database is populated using genetic records that may be obtained from publicly available source databases. For example, in some embodiments human genetic data may be obtained through the Internet using the URL (Uniform Resource Locator) “ftp.ncbi.nih.gov/refseq/H_sapiens/mRNA_Prot/hs_fna.gz” or the URL mgc.nci.nih.gov/. Genetic data for baker’s yeast may be obtained using the URL “genome-ftp.stanford.edu/pub/yeast/data_download/sequence/genomic_sequence/orf_dna.” Genetic data for *E. coli* may be obtained from the URL “www.genome.wisc.edu/sequencing/k12.htm.” Genetic data for *C. elegans* may be obtained using the URL “ftp.wormbase.org/pub/wormbase/confirmed_genes_current.gz”. Genetic data for *Arabidopsis* may be obtained using the URL “tairpub:tairpub@ftp.arabidopsis.org/home/tair/Sequences/blast_datasets/file=ATH1.cds.” It should be noted that no embodiment of the invention is limited to any particular source for the genetic data, and that many publicly and privately available sources may be utilized. In one embodiment, the genetic records represent at least 10% or more, e.g., 25%, 50% or more, of the open reading frames in the genome of a selected organism.

The data format for the source data may be different from the format desired for the genetic database. In some embodiments, the source data is converted to a common format for storage in the genetic database.

A query is issued to search for a subset of records in the genetic database
5 that have at least one recognition site for a predetermined restriction enzyme or for a set of predetermined restriction enzymes (block 304). In one embodiment, one or more predetermined restriction enzymes have a 6, 7 or 8 bp recognition site, e.g., a set may include a predetermined restriction enzyme with a 7 bp recognition site and another with a 8 bp recognition site. However, the present invention is not limited
10 to any particular number of restriction enzymes included in the set or to a particular number of bp in the recognition site for the one or set of predetermined restriction enzymes. The resulting subset of records may be stored in a temporary table, in a separate results table, or in a separate database.

In some embodiments, the resulting subset of genetic records is filtered to
15 exclude records that may lead to erroneous, skewed, or non-useful results (block 306) or include records with selected characteristics. For example, it has been found that very long sequences in excess of 21,000 bp, a size likely to represent one of the largest open reading frames, typically lead to erroneous, skewed or non-useful results. Other filtering characteristics may also be used and are within the scope of
20 the present invention. Examples of such filtering characteristics include filtering for (to exclude or include) a certain GC content, the presence or absence of introns, specific amino acid compositions in the predicted translation product of the open reading frames, similarity to known genes in specific gene families, a particular isoelectric point of predicted protein products of the open reading frames, and/or
25 predicted membrane spanning proteins in the open reading frames. It should be noted that filtering may occur at any point in the method. For example, the records may be filtered prior to populating the genetic database, or as part of the query to create the subset of records at block 2204.

Next, a set of one or more statistics may be obtained by issuing one or more
30 queries on the subset of records having at least one restriction enzyme recognition site (block 308). In some embodiments, the queries comprise pattern matching

queries. The pattern may be specified in any of a number of ways known in the art. For example, wildcard characters may be used to specify one or more positions in the pattern, or regular expressions may be used to specify the pattern. The present invention is not limited to any particular form for specifying a pattern.

- 5 Additionally, the pattern may be submitted as part of a query to a database engine, or the pattern matching may be executed by a program such as a Visual Basic program on records obtained by a query.

10 In some embodiments, the number of records having particular restriction enzyme recognition sites is determined and reported (block 310). In some embodiments, in order to be included in the statistics, each record contains recognition sites for all of a predetermined set of restriction enzymes in order to be analyzed.

15 In alternative embodiments, the number of restriction enzyme target sites occurring in a record is determined and reported (block 312). In some of these alternative embodiments, the record contains recognition sites for all of a predetermined set of restriction enzymes in order to be analyzed.

20 In further alternative embodiments, statistics regarding the bases at ambiguous positions recognized or cleaved by hapaxomeric restriction enzymes are determined and reported (block 314). The statistics are desirable for determining the distribution of bases in the ambiguous positions of those restriction enzymes. Two examples of such ambiguity are the presence of N's in sites recognized or cleaved by *Sfi*I and *Sap*I as illustrated in Figure 1. In these alternative
25 embodiments, the identity of any ambiguous bases in the recognition site(s) or bases between the recognition site(s) and the actual cleavage site(s) of some or all of the predetermined restriction enzymes are determined and reported along with one or more statistics on the identity of these bases.

Figures 4-5 provide the frequency for various restriction enzyme recognition sites in a variety of organisms determined by the method described herein.

III. Vectors of the Invention

Donor or recipient vectors are used to transfer a DNA sequence of interest, e.g., one in a library, e.g., in a cDNA library, in another vector, e.g., an expression vector, or one obtained from an isolated fragment, e.g., a PCR fragment, which
5 DNA sequence of interest is flanked by desirable restriction enzyme recognition sites, to another vector (an acceptor vector) to generate a recipient (expression) vector, e.g., one useful for expression of the DNA sequence of interest. The presence and position of desirable restriction enzyme recognition sites in the acceptor vector and those flanking the DNA sequence of interest permits the rapid
10 subcloning or insertion of the DNA sequence of interest into the acceptor vector in an oriented manner.

The acceptor vector may include sequences 5' and/or 3' to the desirable restriction enzyme recognition sites which encode a peptide or polypeptide (fusion partner), which sequences, when operably linked to the DNA sequence of interest
15 and expressed in a cell, cell lysate or *in vitro* transcription/translation system, yield a fusion protein. Such a peptide or polypeptide may be located at either the N- or C-terminus of the fusion protein. Alternatively, the fusion protein may contain a peptide or polypeptide at both the N- and C-terminus, and each peptide or polypeptide may be different. Alternatively, the DNA sequence of interest may
20 itself encode a fusion protein and, once combined with the acceptor vector, result in a recipient vector which encodes a recombinant polypeptide which includes one or more additional residues at the N-terminus, C-terminus, or both the N- and C-termini, which residues are encoded by sequences in the acceptor vector, e.g., those encoded by sequences 5' and/or 3' to the desirable restriction enzyme recognition
25 sites. Moreover, one or more amino acid residues may be encoded by the exchange sites generated by the ligation of the ends of the DNA sequence of interest and the acceptor vector.

In one embodiment, the peptide or polypeptide fusion partner is an epitope tag, affinity domain, e.g., a protease recognition site, or enzyme, e.g., thioredoxin or
30 dehalogenase. An epitope tag is a short peptide sequence that is recognized by epitope specific antibodies. A fusion protein comprising an epitope tag can be

simply and easily purified using an antibody bound to a chromatography resin. The presence of the epitope tag further allows the recombinant protein to be detected in subsequent assays, such as Western blots, without having to produce an antibody specific for the recombinant protein itself. Examples of commonly used epitope tags
5 include V5, glutathione-S-transferase (GST), hemagglutinin (HA), FLAG, c-myc, RYIRS, calmodulin binding domain, the peptide Phe-His-His-Thr-Thr, chitin binding domain, and the like.

Affinity domains are generally peptide sequences that can interact with a binding partner, such as one immobilized on a solid support. DNA sequences
10 encoding metal ion affinity sequences, such as those with multiple consecutive single amino acids, e.g., histidine, when fused to the expressed protein, may be used for one-step purification of the recombinant protein by high affinity binding to a resin column, such as nickel sepharose. An endopeptidase recognition sequence can be engineered between the polyamino acid tag and the protein of interest to allow
15 subsequent removal of the leader peptide by digestion with enterokinase, and other proteases. Sequences encoding peptides or proteins, such as the chitin binding domain (which binds to chitin), GST (which binds to glutathione), biotin (which binds to avidin and streptavidin), maltose binding protein (MBP), a portion of staphylococcal protein A (SPA), a polyhistidine tract (HIS_n), and the like, can also
20 be used for facilitating purification of the protein of interest. The affinity domain can be separated from the protein of interest by methods well known in the art, including the use of inteins (protein self-splicing elements, Chong et al., Gene, 192:271 (1997)). In one embodiment, sequences for more than one fusion partner can be linked to sequences for a peptide or polypeptide of interest, e.g., an affinity
25 domain is linked to a protease cleavage recognition site which is linked to a polypeptide of interest.

To prepare expression vectors intended to generate defined fusions at the 5' end of an open reading frame (e.g., the acceptor vector does not contain sequences 5' of the exchange site that encode a peptide or protein for fusion), a desired
30 restriction enzyme recognition site is placed at the desired start of transcription in the vector. Care is taken to avoid introducing an ATG or start codon upstream of

the exchange site that might initiate translation inappropriately. For instance, fusion of an overhang generated by *SgfI* digestion of an acceptor vector with a compatible overhang which is 5' to a start codon for an open reading frame in a DNA fragment can yield a recombinant vector containing a *de novo* start site for that open reading

5 frame. Sequences from the acceptor vector which are present in the recombinant vector include sequences 5' to the overhang generated by *SgfI* digestion, which optionally include a suitably positioned RBS. Optionally, sequences at the 5' end of the open reading frame include a Kozak sequence or a portion thereof which, when present in mRNA, is capable of binding the small subunit of a eukaryotic ribosome.

10 To prepare expression vectors intended to generate a fusion protein by fusing a vector encoded peptide or protein located at the N-terminus of a fusion protein to a DNA sequence of interest (i.e., a translational fusion), the restriction enzyme recognition site is positioned in the correct reading frame such that 1) an open reading frame is maintained through the restriction enzyme recognition site on

15 the acceptor vector and 2) the reading frame in the restriction enzyme recognition site on the acceptor vector is in frame with the reading frame found on the restriction enzyme recognition site contained within the donor vector. In addition, the appropriate restriction enzyme recognition site on the acceptor vector is designed to avoid the introduction of in-frame stop codons. The DNA sequence of

20 interest contained within the donor vector is thus cloned in a particular reading frame in the acceptor vector so as to facilitate the creation of the desired N-terminal fusion protein. For example, fusion of *SgfI* sites at the 5' end of a DNA sequence of interest and 3' end of the acceptor vector can provide read through sequences.

Similarly, to prepare expression vectors intended to generate a fusion protein

25 by fusing a vector encoded peptide or protein located at the C-terminus of a fusion protein and a DNA sequence of interest, the restriction enzyme recognition site is positioned in the correct reading frame such that 1) an open reading frame is maintained through the restriction enzyme recognition site on the acceptor vector and 2) the reading frame in the restriction enzyme recognition site on the acceptor

30 vector is in frame with the reading frame found on the restriction enzyme recognition site contained within the donor vector, i.e., a site which flanks the DNA

sequence of interest at the 3' end. The DNA sequence of interest contained within the donor vector can thus be cloned in a particular reading frame so as to facilitate the creation of the desired C-terminal fusion protein. For instance, fusion of a *PmeI* site with a *EcoRV* or *BalI* site can yield a C-terminal fusion with at least 2 amino acids added at the C-terminus, while fusion of two *PmeI* sites or a *PmeI* site and a *DraI* site can yield a C-terminal fusion with a single amino acid added at the C-terminus.

In one embodiment, the expression vector encodes a protein with multiple fusion partners, e.g., an affinity tag for purification and a protease cleavage site fused to a protein of interest.

Use of the cloning system herein makes it possible to bring the protein sequence to be expressed in close proximity to the N-terminal and/or C-terminal fusion partner. A particular advantage is that it is possible to select the reading frame. This makes it possible not only to exactly position the DNA sequence of interest but also to define the ends of the fusion gene.

The vectors employed in the practice of the invention also contain one or more nucleic acid sequences that generally have some function in the replication, maintenance or integrity of the vector, e.g., origins of replication, as well as one or more selectable marker genes. Replication origins are unique DNA segments that contain multiple short repeated sequences that are recognized by multimeric origin-binding proteins and which play a key role in assembling DNA replication enzymes at the origin site. Suitable origins of replication for use in expression vectors employed herein include *E. coli* oriC, colE1 plasmid origin, 2 μ and ARS (both useful in yeast systems), sf1, SV40 EBV oriP (useful in mammalian systems), p15 or those found in pSC101 and the like.

Selection marker sequences are valuable elements in vectors as they provide a means to select for or against growth of cells which have been successfully transformed with a vector containing the selection marker sequence and express the marker. Such markers are generally of two types: drug resistance and auxotrophic. A drug resistance marker enables cells to detoxify an exogenously added drug that would otherwise kill the cell. An auxotrophic marker allows cells to synthesize an

essential component (usually an amino acid) while grown in media which lacks that essential component.

A wide variety of selectable marker genes are available (see, for example, Kaufman, Meth. Enzymol., 185:487 (1990); Kaufman, Meth. Enzymol., 185:537 (1990); Srivastava and Schlessinger, Gene, 103:53 (1991); Romanos et al., in DNA Cloning 2: Expression Systems, 2nd Edition, pages 123-167 (IRL Press 1995); Markie, Methods Mol. Biol., 54:359 (1996); Pfeifer et al., Gene, 188:183 (1997); Tucker and Burke, Gene, 199:25 (1997); Hashida-Okado et al., FEBS Letters, 425:117 (1998)). Common selectable marker gene sequences include those for resistance to antibiotics such as ampicillin, tetracycline, kanamycin, bleomycin, streptomycin, hygromycin, neomycin, Zeocin™, and the like. Selectable auxotrophic gene sequences include, for example, hisD, which allows growth in histidine free media in the presence of histidinol.

Suitable selectable marker genes include a bleomycin-resistance gene, a metallothionein gene, a hygromycin B-phosphotransferase gene, the AURI gene, an adenosine deaminase gene, an aminoglycoside phosphotransferase gene, a dihydrofolate reductase gene, a thymidine kinase gene, a xanthine-guanine phosphoribosyltransferase gene, and the like.

An alternate approach is to use a selectable marker gene that encodes a mutated enzyme that is less active than the corresponding wild-type enzyme. As an illustration, Munir et al., Protein Eng., 7:83 (1994), describe the design of mutant thymidine kinase enzymes with decreased activity (also see Liu and Summers, Virology, 163:638 (1988); Mendel et al., Antimicrob. Agents Chemother., 39:2120 (1995)). Low activity mutants have also been described for adenosine deaminase and dihydrofolate reductase (see, for example, Prendergast et al., Biochemistry, 27:3664 (1988); Jiang et al., Hum. Mol. Genet., 6:2271 (1997); Ercikan-Abali et al., Mol. Pharmacol., 49:430 (1996)).

Another type of marker gene is a gene that produces a readily detectable protein, such as green fluorescent protein, red fluorescent protein, an enzyme (e.g., placental alkaline phosphatase, beta-galactosidase, beta-lactamase, or luciferase), or a cell surface protein that can be detected with an antibody (e.g. CD4, CD8, Class I

major histocompatibility complex (MHC) protein, etc.). The expression products of such selectable marker genes can be used to sort transfected cells from untransfected cells by such standard means, e.g., FACS sorting or magnetic bead separation technology.

5 Metallothionein genes encode proteins that have a high affinity for toxic metals, such as cadmium, zinc, and copper (Beach and Palmiter, Proc. Nat'l Acad. Sci. USA, 78:2110 (1981); Huang et al., EXS, 52:439 (1987); Czaja et al., J. Cell. Physiol., 147:434 (1991)). Accordingly, metallothionein genes provide suitable titratable markers for the methods described herein.

10 In one embodiment, the acceptor vector includes a counterselectable gene flanked by desirable restriction enzyme sites. Preferred genes in this regard include but are not limited to lethal genes, such as those which are inducible with low to no constitutive activity (and preferably with some immunity factor), e.g., genes such as *bar* (*barstar*), those encoding a restriction enzyme (a gene encoding a corresponding
15 methylase), or those encoding nuclease colicins, e.g., E9 DNase, and colicin RNases and tRNases, or gyrase A, as well as *MazF*(*ChpAK*), *Doc* (*Phd*), *ParE*, *PasB*, *StbOrf2*, *HigB*, *z*, *RelE*, *Txe*, *YeoB*, *SacB*, *KilA*, *KorA*, *KorB*, *Kid* (*Kis*), *PemK* (*PemI*), *Hok* (*Sok*), *Dcc* (*Pno*), *CcdB* (*CcdA*), F' plasmid, and the like.

 Other selection approaches include the use of regulated transcriptional
20 modulators, e.g., a tetracycline inducible or repressible system (see, for instance, WO 96/01313).

 The acceptor vectors employed in the practice of the invention also contain one or more nucleic acid sequences that have some function in the expression of a protein, i.e., transcriptional regulatory sequences, for instance, inducible or
25 repressible control sequences such as promoter or enhancer sequences.

 Promoter-enhancer sequences are DNA sequences to which RNA polymerase binds and initiates transcription. The promoter determines the polarity of the transcript by specifying which strand will be transcribed. Bacterial promoters consist of consensus sequences, -35 and -10 nucleotides relative to the
30 transcriptional start, which are bound by a specific sigma factor and RNA polymerase. Eukaryotic promoters are more complex. Most promoters utilized in

vectors are transcribed by RNA polymerase II. General transcription factors (GTFs) first bind specific sequences near the start and then recruit the binding of RNA polymerase II. In addition to these minimal promoter elements, small sequence elements are recognized specifically by modular DNA-binding/trans-activating proteins (e.g., AP-1, SP-1) that regulate the activity of a given promoter. Viral promoters serve the same function as bacterial or eukaryotic promoters and either provide a specific RNA polymerase in *trans* (bacteriophage T7) or recruit cellular factors and RNA polymerase (SV40, RSV, CMV). Viral promoters may be preferred as they are generally particularly strong promoters.

- 10 Promoters may be, furthermore, either constitutive or regulatable (i.e., inducible or derepressible). Inducible elements are DNA sequence elements which act in conjunction with promoters and bind either repressors (e.g., *lacO/LAC I* repressor system in *E. coli*) or inducers (e.g., *Gall/GAL4* inducer system in yeast *rhaBAD/rhamnose* in *E. coli*). In either case, transcription is virtually "shut off" until the promoter is derepressed or induced, at which point transcription is "turned-on".

- Examples of constitutive promoters include the *int* promoter of bacteriophage λ , the *bla* promoter of the β -lactamase gene sequence of pBR322, the CAT promoter of the chloramphenicol acetyl transferase gene sequence of pPR325, and the like. Examples of inducible prokaryotic promoters include the major right and left promoters of bacteriophage (P_L and P_R), the *trp*, *recA*, *lacZ*, *lacI*, *araC* and *gal* promoters of *E. coli*, the α -amylase (Ulmanen et al., J. Bacteriol., 162:176 (1985), the *araBAD* promoter, the *rhaBAD* promoter, and the sigma-28-specific promoters of *B. subtilis* (Gilman et al., Gene Sequence, 32:11 (1984), the promoters of the bacteriophages of Bacillus (Gryczan, In: The Molecular Biology of the Bacilli, Academic Press, Inc., NY, 1982), Streptomyces promoters (Ward et al., Mol. Gen. Genet., 203:468 (1986), Pichia promoters (U.S. Pat. Nos. 4,855,231 and 4,808,537), and the like. Exemplary prokaryotic promoters are reviewed by Glick (J. Ind. Microbiol., 1:277 (1987); Cenatiempo (Biochimie, 68:505 (1986); and Gottesman (Ann. Rev. Genet., 18:415 (1984). In one embodiment, the promoter is a T7 promoter or a SP6 promoter.

Preferred eukaryotic promoters include, for example, the promoter of the mouse metallothionein I gene sequence (Hamer et al., J. Mol. Appl. Gen., 1:273 (1982); the TK promoter of Herpes virus (McKnight, Cell, 31:355 (1982); the SV40 early promoter (Benoist et al., Nature (London), 290:304 (1981); the yeast GalI gene sequence promoter (Johnston et al., Proc. Natl. Acad. Sci. (USA), 79:6971 (1982); Silver et al., Proc. Natl. Acad. Sci. (USA), 81:5951 (1984), a baculovirus promoter, the CMV promoter, the EF-1 promoter, Ecdysone-responsive promoter(s), tetracycline-responsive promoter, and the like.

Suitable prokaryotic vectors include plasmids such as those capable of replication in *E. coli* (for example, pBR322, ColE1, pSC101, PACYC 184, itVX, pRSET, pBAD (Invitrogen, Carlsbad, Calif.), and the like). Such plasmids are disclosed by Sambrook (cf. *Molecular Cloning: A Laboratory Manual*, second edition, edited by Sambrook, Fritsch, & Maniatis, Cold Spring Harbor Laboratory, 1989). *Bacillus* plasmids include pC194, pC221, pT127, and the like, and are disclosed by Gryczan (In: *The Molecular Biology of the Bacilli*, supra, pp. 307-329). Suitable *Streptomyces* plasmids include pIJ101 (Kendall et al., J. Bacteriol., 169:4177 (1987), and streptomyces bacteriophages such as .phi.C31 (Chater et al., In: *Sixth International Symposium on Actinomycetales Biology*, Akademiai Kiado, Budapest, Hungary, pp. 45-54, 1986). *Pseudomonas* plasmids are reviewed by John et al. (Rev. Infect. Dis., 8:693 (1986), and Izaki (Jpn. J. Bacteriol., 33:729 (1978). In one embodiment, the vector backbone for an acceptor vector for expression of linked sequences in *E. coli* includes an *amp*^R gene, T7 transcriptional regulatory elements, and sequences for producing a fusion protein such as a GST, thioredoxin or dehalogenase fusion with a protein of interest.

Suitable eukaryotic plasmids include, for example, BPV, EBV, vaccinia, SV40, 2-micron circle, pCI-neo, pcDNA3.1, pcDNA3.1/GS, pYES2/GS, pMT, pIND, pIND(Sp1), pVgRXR (Invitrogen), and the like, or their derivatives. Such plasmids are well known in the art (Botstein et al., Miami Wntr. Symp., 19:265 (1982); Broach, In: *The Molecular Biology of the Yeast Saccharomyces: Life Cycle and Inheritance*, Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y. pp. 445-470, 1981; Broach, Cell, 28:203 (1982); Dilon et al., J. Clin. Hematol. Oncol.,

10:39 (1980); Maniatis, In: Cell Biology: A Comprehensive Treatise, Vol. 3, Gene Sequence Expression, Academic Press, NY, pp. 563-608, 1980. In one embodiment, the vector backbone for an acceptor vector for expression of linked sequences in mammalian cells or an *in vitro* eukaryotic transcription/translation reaction is
5 pCMVTnT (Promega Corp.), and sequences for producing a fusion protein such as a GST or dehalogenase fusion with a protein of interest.

Promoters/plasmid combinations are employed with suitable host cells, e.g., prokaryotic cells, such as *E. coli*, *Streptomyces*, *Pseudomonas* and *Bacillus*, or eukaryotic cells, such as yeast, e.g., *Picchia*, *Saccharomyces* or
10 *Schizosaccharomyces*, insect cells, avian cells, plant cells, or mammalian cells, e.g., human, simian, porcine, ovine, rodent, bovine, equine, caprine, canine or feline cells, as well as lysates thereof, e.g., TNT, wheat germ lysates or S30 lysates.

In one embodiment, the host cell is a recombinant cell, e.g., a recombinant prokaryotic cell. In one embodiment, the recombinant host cell is deficient in one or
15 more genes in an inducible pathway, e.g., a sugar pathway such as the rhamnose catabolic pathway, and comprises a recombinant DNA comprising an inducible promoter for the one or more genes operably linked to an open reading frame for a heterologous RNA polymerase. The recombinant host cell or a lysate there, or an *in vitro* transcription/translation mixture supplemented with the heterologous RNA
20 polymerase, is contacted with a vector of the invention comprising a promoter for the heterologous RNA polymerase operably linked to a DNA sequence of interest. In one embodiment, the recombinant host cell is a recombinant *E. coli* cell that is deficient in rhamnose catabolism and comprises a *rhaBAD* promoter operably linked to a T7 RNA polymerase open reading frame. In the absence of rhamnose,
25 such a cell has no or low levels of T7 RNA polymerase and so is particularly useful to clone toxic genes.

In another embodiment, the recombinant host cell expresses an immunity factor for a gene product that is lethal to the cell. The immunity factor is preferably expressed from a constitutive promoter. An expression vector encoding the lethal
30 gene product may be introduced to the recombinant cell and the transformed cell propagated. In one embodiment the gene product is barnase which has been

modified by deleting sequences for the secretory segment (signal peptide) and optionally adding a ATG in place of the last codon for the secretory sequence.

IV. Use of DNA Binding Proteins to Protect Restriction Enzyme Sites

In the process of introducing a DNA sequence of interest to a donor vector, or from a donor vector to an acceptor vector, restriction enzyme sites which flank the DNA sequence of interest, i.e., those useful in cloning, may also be present in either the DNA sequence of interest or vector sequences. To protect sites containing a particular restriction enzyme site from cleavage by the corresponding enzyme, DNA binding proteins and methylation may be employed. For instance, the process of protecting a restriction site with RecA (RecA cleavage and production) is more reproducible, provides better yields and is less cumbersome than partial restriction digests. Other means of protecting a restriction site include using repressor proteins, eukaryotic transcription factors, *E.coli* host integration factor or oligonucleotides capable of forming a triple helix structure, however, the specificity of protection using RecA is entirely from the synthetic single-stranded DNA. In the presence of a nonhydrolyzable ATP analog such as ATP[gamma-S], the RecA protein nonspecifically binds to single-stranded DNA (ssDNA) (approximately one RecA monomer per three nucleotides) to form a structure called a presynaptic filament. This RecA-coated oligonucleotide then anneals with homologous duplex DNA to form a stable triplex DNA-protein complex. The presynaptic filament represents a useful molecular research tool in that: i) the sequence and length of the ssDNA added to the reaction determines the site and span of the presynaptic filament and ii) the presynaptic filament protects the DNA at the hybridization site from modification by DNA methylases and restriction enzymes. These features enable RecA protein-mediated DNA complexes to add a new level of specificity to molecular biology applications that require DNA cleavage at predetermined sites, such as genomic mapping and the subcloning of DNA fragments. Compared to PCR methods, the use of a DNA binding protein is quicker and does not introduce mutations arising from multiple cycles of *in vitro* amplification.

The general protocols include protecting a restriction site from methylation, making it unique for restriction enzyme cleavage (RecA cleavage), and protecting a

restriction site from digestion (RecA protection). The RecA cleavage protocol is based on the RecA Achilles' cleavage procedure of Koob et al. (Science, 241, 1084 (1988)), Koob et al. (Gene, 74, 165 (1988)), and Koob et al. (Nucle. Acids Res., 20, 5831 (1992)). Additionally, RecA cleavage is useful for generating restriction

5 fragments for subcloning when the desired restriction site is repeated several times within the fragment. However, if only one or two restriction sites are repeated within the desired fragment, RecA protection is preferred. Based on fluorometric analysis of the RecA products after electrophoresis, these two protocols routinely resulted in 70% to 80% protection when a single site was protected. This technique

10 also can be used for DNA embedded in agarose plugs.

Table 3

Oligonucleotides: Prepared by user to be specific for the intended protected site. Diluted to 160 ng/μl.	Methylase: In theory, any restriction enzyme/methylase pair could be used. In these protocols, 35 μ/μl of <i>EcoRI</i> methylase was used.
RecA: 1-3 mg/ml	Restriction enzyme: In these protocols, 12 μ/μl <i>EcoRI</i> was used.
SAM: 1.6 mM S-adenosyl methionine. Prepared immediately before use from a 32 mM stock by dilution with ice-cold 5 mM sulfuric acid.	Buffer A: 250 mM Tris-acetate (pH 7.5 at 25°C), 1 mM magnesium acetate.
ATP [gamma-S]: Aliquots of a 10 mM solution (in water) are stored at -70°C.	Buffer B: 166 mM Tris-acetate (pH 7.5 at 25°C), 37 mM magnesium acetate, 100 mM DTT.
80 mM magnesium acetate.	250 mM potassium acetate.
Restriction Enzyme Buffer H (Promega)	

A. RecA Cleavage or Protection Reactions

15 The RecA Concentration

To maximize the specificity and efficiency of RecA protection, it may be necessary to manipulate the oligonucleotide:RecA ratio: a concentration of 6.25 μg RecA in a 10 μl reaction works well.

The Oligonucleotide Concentration

The molar stoichiometry (in terms of moles of nucleotides to moles of RecA protein) of the binding of the oligonucleotide to RecA is 3:1. In other words, one RecA protein binds every three nucleotides of any single-stranded DNA. This ratio is independent of oligonucleotide size and corresponds to 160 ng of oligonucleotide per 6.25 µg RecA. A titration series of 40-280 ng in 40 ng increments is useful to determine the optimal concentration of oligonucleotide to use with the RecA. If nonspecific protection is a problem, then 160 ng of oligo(dT) can be added to the reaction after the addition of ATP[gamma-S].

Design of Oligonucleotide

10 An oligonucleotide of 30 to 36 bases in length is recommended for both RecA cleavage and RecA protection in solution. The protected site was located in the middle of the 30 base oligonucleotide used throughout the development of this protocol (see also RecA Cleavage and Protection for Genomic Mapping and Subcloning, from Promega Notes Magazine #50).

15 Buffer

It may be necessary to adjust the salt concentration to improve the activity of the enzyme after methylation. Acetate salts appear to be less destabilizing to the RecA triplex than chloride salts, and thus potassium acetate rather than potassium or sodium chloride may be employed.

20 Subcloning the Products of RecA Cleavage

Because the products of a RecA cleavage reaction are methylated, low transformation frequencies may arise from incompatibilities with the host's restriction/modification system. If transformation efficiencies are low, compare the genotype of the host to the known methylation-induced restriction systems to determine if this is the cause.

IV. Exemplary Vector Systems

In one embodiment, at least one of the restriction enzyme sites in the donor vector and/or flanking the DNA sequence of interest is for a restriction enzyme with a degenerate recognition sequence, e.g., *Sfi*I is a restriction enzyme with a
30 degenerate recognition sequence that recognizes an interrupted palindromic sequence (Figure 6). To employ restriction enzymes that recognize an interrupted

palindromic sequence and generate single-strand DNA overhangs for use in directional cloning, at least two unique sites for that restriction enzyme and/or unique site(s) for a different restriction enzyme that generates non-self complementary single-strand DNA overhangs that are complementary with the overhangs generated by the first restriction enzyme are employed. Other methods may be used to enhance the frequency of desired vectors, e.g., the use of methylation, and/or selectable and counterselectable genes.

Figure 7 shows a schematic of the use of donor and acceptor vectors having restriction enzyme sites for a restriction enzyme which recognizes an interrupted palindrome (enzyme I; the unique sequences are indicated by A and B, their complements by A' and B', respectively, and the palindromic sequences by boxes). The donor vector has a drug resistance gene 1 and a DNA sequence of interest (light grey box) flanked by one or more restriction enzyme sites for the restriction enzyme which recognizes an interrupted palindrome. The acceptor vector has a different drug resistance gene (drug resistance gene 2) and, after digestion with a restriction enzyme with a degenerate recognition sequence, has non-self complementary single-strand DNA overhangs A' and B' which are, respectively, complementary with the non-self complementary single-strand DNA overhangs present after digestion of the donor vector with enzyme I. Thus, after digestion of the donor vector with enzyme I and in the presence of the linearized acceptor vector and ligase, the linearized DNA sequence of interest is joined in an oriented manner to the acceptor vector, to yield a recipient vector. In Figure 7A, one half site of the restriction site for enzyme I is present at each end of the DNA sequence of interest in the recipient vector. If the ligation regenerates the restriction site, then there is a competing back reaction (Figure 7B). In Figure 7C, a counterselectable gene (a lethal gene) is employed in the acceptor vector so that cells with the recipient vector rather than the acceptor vector can be readily identified.

Figure 8 shows one method by which a DNA sequence of interest is modified to contain restriction enzyme sites for a restriction enzyme with a degenerate recognition sequence. Oligonucleotides having unique degenerate sequences for the restriction enzyme at the 5' end, and sequences complementary to

one of the strands of the DNA sequence of interest at the 3' end, are employed in an amplification reaction. Those unique sequences are also present in a vector containing a drug resistance gene. The amplified fragment and the vector are digested with the restriction enzyme and ligase added to yield a donor vector of the invention. If the sites are recognized by restriction enzymes which are sensitive to the methylation state of DNA, e.g., at Dcm sites or using a methylase for *Sfi*I, methylation may minimize the back reaction. The donor vector is then digested with a restriction enzyme(s) having degenerate recognition sequences and which releases the DNA sequence of interest, and mixed with an acceptor vector having complementary single-strand DNA overhangs generated by, for example, a different enzyme with a degenerate recognition sequence that generates non-self complementary single-strand DNA overhangs.

Figures 9A-B show another approach to preparing a donor vector of the invention. A DNA sequence of interest is modified to contain restriction enzyme sites for a restriction enzyme with a degenerate recognition sequence. Oligonucleotides having unique degenerate sequences for the restriction enzyme at the 5' end, and sequences complementary to one of the strands of the DNA sequence of interest at the 3' end, are employed in an amplification reaction. The DNA sequence of interest may include internal sites for that restriction enzyme. To protect those internal sites from digestion, they are methylated, while the flanking sites at the ends of the amplified fragment remain unmethylated and therefore sensitive to digestion. To accomplish this, oligonucleotides complementary to the sites which are to remain unmethylated and a DNA binding protein such as RecA are added to the amplified fragment. The internal sites are then methylated with an appropriate methylase. A column may be employed to remove the oligonucleotide-DNA binding protein complexes from the amplified fragment. The sites which were added to the ends of the DNA sequence of interest, once digested, yield non-self complementary single-strand DNA overhangs. Complementary overhangs may be generated in a vector by digestion with a selected restriction enzyme with degenerate recognition sites, which enzyme may be different than the enzyme employed to digest the amplified fragment. The amplified fragment and the vector

are then digested with the one or more restriction enzymes, and the resulting linear fragments ligated to form a donor vector containing a drug resistance gene and the DNA sequence of interest flanked by sites generated by the joining of the complementary single-strand DNA overhangs, which sites are recognized by one or more restriction enzymes with a degenerate recognition sequence, e.g., the enzyme employed to digest the amplified fragment.

Figures 10A-B illustrate an approach to prepare a recipient vector of the invention. In this embodiment, a donor vector comprises a drug resistance gene and a DNA sequence of interest flanked by restriction enzyme sites for an enzyme with a degenerate recognition sequence and containing one or more of those sites internally. To protect those internal sites from digestion, they are methylated. To ensure that the flanking sites remain unmethylated and thus sensitive to digestion, oligonucleotides complementary to the sites which are to remain unmethylated and a DNA binding protein are added to a donor vector. The site(s) for the restriction enzyme which are not bound by the oligonucleotide/DNA binding protein is/are then methylated with an appropriate methylase. A column may be employed to remove the oligonucleotide-DNA binding protein complexes from the donor vector. The donor vector is then added to an acceptor vector having at least two recognition sites for a restriction enzyme with a degenerate recognition sequence, which restriction enzyme produces non-self complementary single-strand DNA overhangs which are complementary to the overhangs generated by digestion of the donor vector with a restriction enzyme that cleaves the unmethylated sites. The acceptor vector preferably comprises a drug resistance gene which is different than the drug resistance gene in the donor vector. In one embodiment, the restriction enzyme used to digest the acceptor vector may be different than the restriction enzyme employed to digest the donor vector. Subsequent ligation of the linearized DNA fragments obtained by digestion of the donor and acceptor vectors yields a recipient vector.

In one embodiment, the restriction enzyme used to linearize the donor vector and the acceptor vector are the same, for instance, the donor vector has unique *Sfi*I sites flanking the DNA sequence of interest, which sites, once digested with *Sfi*I,

yield non-self complementary single-strand DNA overhangs that are complementary with the single-strand DNA overhangs generated after digestion of the acceptor vector with *Sfi*I. In another embodiment, the donor vector has unique *Bgl*II sites flanking the DNA sequence of interest which sites, once digested with *Bgl*II, yield
5 non-self complementary single-strand DNA overhangs that are complementary with the single-strand DNA overhangs generated after digestion of the acceptor vector with *Bgl*II. In another embodiment, the restriction enzyme with a degenerate recognition sequence used to linearize the donor vector and the acceptor vector is different, for instance, the donor vector has unique *Sfi*I sites flanking the DNA
10 sequence of interest which sites, once digested with *Sfi*I, yield non-self complementary single-strand DNA overhangs that are complementary with the single-strand DNA overhangs generated after digestion of the acceptor vector with *Bgl*II. Restriction enzymes useful with *Sfi*I in preparing donor and acceptor vectors are shown in Figure 11. Methylases for *Sfi*I and/or *Bgl*II may be obtained by well-
15 known methods, see, e.g., U.S. Patent Nos. 5,179,015, 5,200,333, and 5,320,957. For instance, the preparation of recombinant *Bgl*II and its corresponding methylase is disclosed in U.S. Patent No. 5,366,882. The preparation of recombinant *Sfi*I and a corresponding methylase is provided in U.S. Patent No. 5,637,476. Other methylases useful with vectors containing *Sfi*I recognition sites include the
20 methylase for *Hae*III and *Dcm* methylase.

In another embodiment, at least one of the restriction enzyme sites in the donor vector and/or flanking the DNA sequence of interest is a site for a type IIS enzyme, e.g., *Sap*I. Figure 12 illustrates the preparation of a recipient vector of the invention from a donor vector and an acceptor vector using vectors with recognition
25 sites for type IIS restriction enzymes. To employ sites for type IIS restriction enzymes in directional cloning, at least two unique sites for that restriction enzyme and/or unique site(s) for a different restriction enzyme that generates non-self complementary single-strand DNA overhangs that are complementary with the overhangs generated by the first restriction enzyme are selected. Methylation may
30 be employed to increase the frequency of desired vectors, as well as the use of selectable and counters selectable genes.

In one embodiment, the restriction enzyme used to linearize the donor vector and the acceptor vector are the same, for instance, the donor vector has unique *SapI* sites flanking the DNA sequence of interest, which sites, once digested with *SapI*, yield non-self complementary single-strand DNA overhangs that are complementary with the single-strand DNA overhangs generated after digestion of the acceptor vector with *SapI*. In another embodiment, the donor vector has unique *EcoRI* sites flanking the DNA sequence of interest which sites, once digested with *EcoRI*, yield non-self complementary single-strand DNA overhangs that are complementary with the single-strand DNA overhangs generated after digestion of the acceptor vector with *EcoRI*. In another embodiment, the restriction enzyme used to linearize the donor vector and the acceptor vector is different, for instance, the donor vector has unique *SapI* sites flanking the DNA sequence of interest, which sites, once digested with *SapI*, yield non-self complementary single-strand DNA overhangs that are complementary with the single-strand DNA overhangs generated after digestion of the acceptor vector with *EcoRI*. The preparation of *SapI* and a corresponding methylase are disclosed U.S. Patent No. 5,663,067.

In contrast to the use of *SfiI* vectors for directional cloning, which yields 12 bases (3 potential codons) at the exchange sites, the use of *SapI* vectors yields 3 bases (1 potential codon) at the exchange sites. Thus, *SapI* vectors are particularly useful in recipient vectors as the protein encoded by the DNA sequence of interest in the recipient vector may include only two additional residues, one at the N-terminus and one at the C-terminus, e.g., a codon for methionine at the N-terminus and a residue at the C-terminus which is frequently found at or near the C-terminus of a plurality of proteins. Accordingly, proteins expressed from *SapI* vectors are very close in composition to their corresponding native protein. Moreover, the overlapping sequences which form the exchange site may be chosen to correspond to codons employed at a certain frequency in a particular organism.

In another embodiment, shown in Figures 14-15, a two enzyme approach is used for directional cloning. For the donor vector, the DNA sequence of interest is flanked by at least two restriction enzymes sites. One of the sites is for a first restriction enzyme which is an infrequent cutter of cDNAs or open reading frames

in at least one species and generates single-strand DNA overhangs while the other site is for a second restriction enzyme that is also an infrequent cutter of cDNAs or open reading frames in at least one species and generates ends that are not complementary to the ends generated by the first restriction enzyme. In one embodiment, the second restriction enzyme generates blunt ends. For instance, a donor vector has a drug resistance gene 1 and a DNA sequence of interest flanked by a restriction enzyme site for an enzyme (enzyme I) that is an infrequent cutter of human cDNAs or open reading frames and generates a single-strand DNA overhang, e.g., *SgfI*, and by a site for a restriction enzyme (enzyme II) that in an infrequent cutter in that same species and generates blunt ends, e.g., *PmeI*. The donor vector which, optionally, is an expression vector, is mixed with an acceptor vector, which has a different drug resistance gene, and at least two restriction enzyme sites, and optionally a counter-selectable gene,. One of the sites in the acceptor vector is for a restriction enzyme (enzyme III) that generates single-strand DNA overhangs which are complementary to those generated by enzyme I, e.g., *PvuI* or *PacI*, and a restriction enzyme site for an enzyme (enzyme IV) which generates ends which can be ligated to the ends generated by enzyme II, e.g., enzyme IV generates blunt ends, for instance, enzyme IV is *PmeI*, *EcoRV*, *BalI*, or *DraI*. After digestion with the enzymes, ligation of the linearized donor and acceptors vectors yields a recipient vector comprising the different drug resistance gene and the DNA sequence of interest which is joined to acceptor vector sequences via ligation of the two pairs of complementary single-strand DNA overhangs, or via ligation of complementary single-strand DNA overhangs and blunt ends.

In one embodiment, a DNA sequence of interest is modified to contain restriction enzyme sites for a restriction enzyme which is an infrequent cutter of cDNAs or open reading frames in at least one species and generates single-strand DNA overhangs (enzyme I) and a restriction enzyme that is an infrequent cutter of cDNAs or open reading frames and generates ends that are not complementary to the ends generated by the first restriction enzyme or blunt ends (enzyme II) (Figure 15). The DNA sequence of interest is mixed with an oligonucleotide having complementary sequences to the site for the infrequent cutter which generates

single-strand DNA overhangs and an oligonucleotide having complementary sequences to the site recognized by the enzyme which is an infrequent cutter and generates ends that are not complementary to the ends generated by the first restriction enzyme, e.g., blunt ends, and the mixture is subjected to an amplification reaction, yielding a DNA fragment. In one embodiment, the second restriction enzyme is a blunt cutter. The sites which were added to the ends of the DNA sequence of interest, once digested, yield a single-strand DNA overhang at each end, or a single-strand DNA overhang at one end and a blunt end at the other. Complementary single-strand DNA overhangs to the overhangs generated by enzyme I, or a complementary single-strand DNA overhang to the overhangs generated to enzyme I and a blunt end, are generated in an acceptor vector with restriction enzymes III and IV, respectively, yielding a linearized acceptor vector. The linearized acceptor vector, which comprises a drug resistance gene, is ligated to the digested DNA fragment, to result in a recipient vector. The recipient vector contains the drug resistance gene of the acceptor vector and the DNA sequence of interest flanked by sites generated by the joining of the complementary single-stranded DNA overhangs at each end, or the complementary single-strand DNA overhangs at one end and the blunt ends at the other. The *SgfI/PmeI* approach can result in a recipient vector which encodes a protein with no additional residues at the N-terminus of the protein, e.g., one positioned 3' to a RBS or Kozak sequence or encoding a fusion protein with an N-terminal or C-terminal fusion of one or more amino acid residues (Figures 16-17 and Table I, which shows enzymes which generate blunt ends and the exchange site created by ligation of a blunt end generated by *PmeI* and a blunt end generated by each of those enzymes).

Enzymes	Recognition Sequence	Stop Codons w/PmeI	Codon(AA) fusion	Isoschizomers
<u>Aha</u> III	TTT^AAA	TAA	None	DraI PauAII SruI
<u>Alu</u> I	AG^CT	No	TCT(S)	MltI
<u>Bal</u> II	TGG^CCA	No	TCC(S)ANN(IMTNKSR)	MisI Mlu3II MluNI MscI
<u>Bfr</u> BI	ATG^CAT	No	TCA(S)TNN(FLSYC)	Msp20I
<u>Bsa</u> AI	YAC^GTR	No	TGT(C)RNN(IMVTANKDESRG)	-
<u>Bsa</u> BI	GATNN^NNATC	TAA, TAG, TGA	TNN(FLSYCW)ATC(I)	BstBAI MspYI PsuAI
<u>Bsr</u> BI	CCGCTC (-3/-3)	No	TCT(S)CNN(LPHQR)	Bse8I BseJI Bsh1365I BsiBI
<u>Btr</u> I	CACGTC (-3/-3)	No	TGT(C)CNN(LPHQR)	BsrBRI MamI
<u>Cac</u> 8I	GCN^NGC	TAG	TNG(LSW)CNN(LPHQR)	AccBSI BstD102I Bst31NI
<u>Cdi</u> I	CATCG (-1/-1)	TGA	TGN(C)	MbiI
<u>Cvi</u> JI	RG^CY	No	TCY(S)	BmgBI
<u>Cvi</u> RI	TG^CA	No	TCA(S)	BstC8I
<u>Eco</u> 47III	AGC^GCT	No	TGC(C)TNN(FLSYC)	-
<u>Eco</u> 78I	GGC^GCC	No	TGC(C)CNN(LPHQR)	CviTI
<u>Eco</u> ICRI	GAG^CTC	No	TCT(S)CNN(LPHQR)	HpyCH4V HpyF44III
<u>Eco</u> RV	GAT^ATC	No	TAT(Y)CNN(LPHQR)	AfeI AitI Aor51HI FunI
<u>Esa</u> BC3I	TC^GA	TGA	None	EgeI Ehel SfoI
				BpuAml Ecl136II Eco53kl
				MxaI
				CeqI Eco32I HjaI HpyCI
				NsiCI
				-
				AccII BceBI BepI Bpu95I
				Bsh1236I Bsp50I Bsp123I
				BstFNI BstUI Bsu1532I BtkI
				Csp68KVI CspKVI FallI
<u>Fnu</u> DII	CG^CG	No	TCG(S)	FauBII MvnI Thal

<u>Enzymes</u>	<u>Recognition Sequence</u>	<u>Stop Codons w/PmeI</u>	<u>Codon(AA) fusion</u>	<u>Isoschizomers</u>
<u>FspAI</u>	RTGC^GCAY	No	TGC(C)AYN(IMT)	-
<u>HaeI</u>	WGG^CCW	No	TCC(S)WNN(IMTNKSRFLYC)	-
				BanAI BclAII BimI9II Bme36II BseQI BshI BshFI Bsp21II BspBRI BspKI BspRI BsuRI BteI CtiI DsaII EsaBC4I FnuDI MchAII MfoAI NgoPII NspLKI Pali Pde133I PfiKI PfoI PlaI SbvI SfaI Sual
<u>HaeIII</u>	GG^CC	No	TCC(S)	
<u>HindIII</u>	GTY^RAC	TAA, TGA	None	HinJCI HincII
<u>HpaI</u>	GTT^AAC	TAA	None	BstEZ359I BstHPI KspAI SsrI
<u>Hpy8I</u>	GTN^NAC	TAA, TGA	TYA(FLS)CNN((LPHQR)	HpyBII
<u>LpnI</u>	RGC^GCV	No	TGC(C)YNN(FLSYCLPHQR)	Bme142I
<u>MivI</u>	GAGTC (5/5)	TAA, TAG, TGA	Any	SchI
<u>MslI</u>	CAYNN^NNRTG	TAA, TAG, TGA	TNN(FLSYCW)RTG(MV)	SmiMI
<u>MstI</u>	TGC^GCA	No	TGC(C)ANN(IMTNKSR)	Acc16I Aosl AviII FdiII FspI NsbI Paml Pun14627I
				CcoI PdiI SauBMKI SauHPI SauLPI SauNI SauSI Stu1777I SspCI
<u>NaeI</u>	GCC^GGC	No	TGG(C)CNN(LPHQR)	AspNI BscBI BspLI PspN4I
<u>NlaIV</u>	GGN^NCC	No	TNC(FSYC)CNN(LPHQR)	Bsp68I MluB2I Sbo13I Spol
<u>NruI</u>	TCG^CGA	No	TCG(S)ANN(IMTNKSR)	MspAII
<u>NspBII</u>	CMG^CKG	No	TCK(S)GNN(VAPEG)	AleI
<u>OliI</u>	CACNN^NNGTG	TAA, TAG, TGA	TNN(FLSYCW)GTG(V)	AcvI BbrPI BcoAI Eco72I PmlI
<u>PmaCI</u>	CAC^GTG	No	TGT(S)GNN(VAPEG)	MssI
<u>PmeI</u>	GTTT^AAAC	TAA	None	

Enzymes	Recognition Sequence	Stop Codons w/PmeI	Codon(AA) fusion	Isoschizomers
<u>Psh</u> AI	GACNN^NNGTC	TAA, TAG, TGA	TNN(FLSYCW)GTC(V)	BoxI BstPAI
<u>Psi</u> I	TTA^TAA	No	TTA(L)ANN(IMTNKSR)	-
<u>Pvu</u> II	CAG^CTG	No	TCT(S)GNN(VADEG)	BavI BavAI BspI BspI53AI BspM39I BspO4I Cfr6I DmaI EclI NmeRI PaeI7kI PunI4627II Pvu84II UbaI53AI UbaM39I
<u>Rsa</u> I	GT^AC	No	TAC(Y)	AfaI HpyBI PlaAI
<u>Sca</u> I	AGT^ACT	No	TAC(Y)TNN(FLSYCW)	AccI13I AssI DpaI Eco255I RflFI
<u>Sci</u> I	CTC^GAG	TGA	None	-
<u>Sma</u> I	CCC^GGG	No	TGG(C)GNN(VADEG)	CfrI4I PaeBI PspALI
<u>Sna</u> BI	TAC^GTA	No	TGT(S)ANN(IMTNKSR)	BstSNI Eco105I
<u>Srf</u> I	GCCC^GGG	No	TGG(C)GCN(A)	-
<u>Ssp</u> I	AAT^ATT	No	TAT(Y)ANN(IMTNKSR)	-
<u>SspD5</u> I	GGTGA (8/8)	TAA, TAG, TGA	Any	-
<u>Stu</u> I	AGG^CCT	No	TCC(S)TNN(FLSYCW)	AatI AspMI Eco147I GdlI PceI Pme55I SarI Sru30DI SseBI Stel
<u>Swa</u> I	ATTT^AAAT	TAA	None	BstRZ246I BstSWI MspSWI SmiI
<u>Xba</u> I	GTA^TAC	No	TTA(L)CNN(LPHQR)	BspM90I BssNAI BstI107I BstBSI BstZ17I
<u>Xmn</u> I	GAANN^NNTTC	TAA, TAG, TGA	TNN(FLSYCW)TTC(F)	Asp700I BbvAI MroXI PdmI
<u>Zra</u> I	GAC^GTC	No	TGT(S)CNN(LPHQR)	-

The *SgfI/PmeI* approach may also be used to introduce two DNA fragments of interest into the same vector (Figures 18-19). For example, a donor vector is obtained or prepared that contains a drug resistance gene 1 and a DNA sequence of interest flanked by a restriction site for a restriction enzyme which is an infrequent cutter of cDNAs or open reading frames in at least one species and generates single-strand DNA overhangs (enzyme I), e.g., *SgfI*, and a site for a restriction enzyme which is an infrequent cutter of cDNAs or open reading frames and generates blunt ends (enzyme II), e.g., *PmeI*. An acceptor vector is prepared or obtained that contains a drug resistant gene 2, a restriction site for a restriction enzyme (enzyme III) which generates single-strand DNA overhangs that are complementary to the overhangs in a donor vector linearized with enzyme I, which restriction enzyme is different than enzyme I, e.g., *PvuI*, and a restriction site for an enzyme which generates blunt ends (enzyme IV), and is different than enzyme II, e.g., *HpaI*. The acceptor vector also includes two additional restriction sites, each of which are 5' or 3' to the DNA sequence of interest in the acceptor vector, one of which is for a restriction enzyme (enzyme V) which generates single-strand DNA overhangs that are complementary to the overhangs generated by enzyme I, which restriction enzyme is different than enzyme I, e.g., *PacI*, and another for a restriction enzyme that generates blunt ends (enzyme VI), which enzyme is different than enzyme II or enzyme IV, e.g., *SwaI*. The donor vector is linearized with enzyme I and enzyme II and ligated to an acceptor vector linearized with enzyme III and enzyme IV, to yield a recipient vector having drug resistance gene 2, the DNA sequence of interest, and sites for restriction enzymes V and VI which are both 5' or 3' to the DNA sequence of interest. A second donor vector having a drug resistance gene and a different DNA sequence of interest flanked by a restriction site for enzyme I and another for enzyme II is digested with enzymes I and II, and mixed with the recipient vector, which is linearized with enzymes V and VI, resulting in a second recipient vector having both DNA fragments of interest. Such a recipient vector is useful to study protein-protein interactions, e.g., in two hybrid or colocalization studies, and is particularly useful in systems in which one protein is not expressed or is only

expressed at low levels in the absence of expression of a binding protein for that protein.

The invention will be further described by the following non-limiting examples.

5

Example I

An ampicillin-sensitive donor vector was prepared which has a green light emitting luciferase gene flanked by *Sfi*I sites which, after digestion, do not yield complementary single-strand DNA overhangs (Figure 20A). An ampicillin resistant
10 acceptor vector was also prepared which has a red light emitting luciferase gene flanked by *Sfi*I sites which, after digestion, do not yield complementary single-strand DNA overhangs but each of which is complementary to one of the single-strand DNA overhangs flanking the green light emitting luciferase gene. These two vectors were digested in T4 DNA ligase buffer with *Sfi*I at 50°C for 1 hour. The
15 reactions were cooled to room temperature, and T4 DNA ligase added. The ligation reaction was conducted at 22°C for 30-60 minutes. A portion of the ligation reaction was subjected to gel electrophoresis, while another portion was used to transform JM109. The transformed cells were placed on nitrocellulose and incubated overnight.

20 The filter was floated on 1 ml 100 mM citrate (pH 5.5) with 1 mM luciferin potassium salt at 40°C. An image was then obtained with a CCD digital camera (Minolta Dimage 7; 4 seconds *f*4.5). The results show that *Sfi*I cuts in ligase buffer, and that the cut ends religate in the presence of T4 DNA ligase (Figure 20B). To improve the number of desired clones, an acceptor vector containing a
25 counterselectable marker may be employed.

Example II

Vectors

The pDONOR-4 CAT vector was utilized as the source for the
30 chloramphenicol acetyl transferase (CAT) reporter gene with its native promoter between the *Sgf*I and *Pme*I sites. pDONOR-4 contains a kanamycin resistance gene

for bacterial selection, and restriction enzyme sites *SgfI* and *PmeI* for directional and flexible cloning.

The pDONOR-6 LacZ vector was utilized as the source for the LacZ reporter gene. pDONOR-6 contains a kanamycin resistance gene for bacterial selection, a T7 bacteriophage promoter, and restriction enzyme sites *SgfI* and *PmeI* for directional and flexible cloning.

The pACCEPT-F vector (Figure 21A) was utilized as the source of the backbone sequence for the reporter genes. pACCEPT-F contains an ampicillin resistance gene for bacterial selection, a T7 bacteriophage promoter, and restriction enzyme sites *SgfI* and *PmeI* for directional and flexible cloning.

Results

The LacZ reporter gene from pDONOR-6 LacZ was transferred to pACCEPT-F in a two step process. First, pDONOR-6 LacZ was digested with the restriction enzymes *SgfI* and *PmeI* in Promega Buffer C with BSA at 37°C for 1 hour to free the LacZ gene from the vector. Following digestion, the restriction enzymes were inactivated by heating the reaction tube to 65°C for 20 minutes. Second, linearized pACCEPT-F, T4 DNA ligase, ATP, DTT and additional Buffer C were added to the reaction tube and ligation was initiated by incubating the tube at 22°C for 1 hour. Following ligation, an aliquot of the reaction was transformed into *E. coli* cells (JM109), and the transformation mixture was plated onto Luria Broth (LB) plates containing ampicillin, X-Gal, and rhamnose. The colonies were visually screened for their ability to utilize X-Gal thereby producing a blue color. Results demonstrated that approximately 90% of the colonies produced a blue color, demonstrating the percent transfer of the LacZ gene from the pDONOR-6 LacZ to the pACCEPT-F vector (percentage was calculated by total # blue colonies/ total# colonies x 100).

The LacZ reporter gene from pDONOR-6 LacZ was also transferred to the pDEST-F in a two step process. First, vectors pDONOR-6 LacZ and pACCEPT-F were digested in one tube with the restriction enzymes *SgfI* and *PmeI* in Promega Buffer C with BSA at 37°C for 1 hour to free the LacZ gene from the vector. Following digestion, the restriction enzymes were inactivated by heating the

reaction tube to 65°C for 20 minutes. Second, T4 DNA ligase, ATP, DTT and additional Buffer C were added to the reaction tube and ligation was initiated by incubating the tube at 22°C for 1 hour. Following ligation, an aliquot of the reaction was transformed into *E. coli* cells (JM109), and the transformation mixture was plated onto LB plates containing ampicillin, X-Gal, and rhamnose. Results demonstrated that approximately 81% of the colonies produced a blue color.

The CAT reporter gene from pDONOR-4 CAT was transferred to the pACCEPT-F in a two step process. First, pDONOR-4 CAT was digested with *SgfI* and *PmeI* in Promega Buffer C with BSA at 37°C for 1 hour to free the CAT gene from the vector. Following digestion, the restriction enzymes were inactivated at 65°C for 20 minutes. Second, linearized pACCEPT-F, T4 DNA ligase, ATP, DTT, and additional Buffer C were added to the reaction tube and ligation was performed at 25°C for 1 hour. Following ligation, an aliquot of the reaction was transformed into *E. coli* JM109 bacterial cells, and the transformation mixture was plated onto LB plates with ampicillin. Of the resultant colonies, 100 were re-plated onto LB plates with chloramphenicol. Colonies which grew on chloramphenicol contained the CAT gene. Transfer efficiency of the CAT gene from the pDONOR-4 CAT to pACCEPT-F vector was determined to be approximately 94% (percentage was calculated by total # CAT resistant colonies/ total# colonies tested x 100).

The CAT reporter gene from pDONOR-4 CAT was transferred to the pACCEPT-F in a one step process. To the reaction tube was added pDONOR-4 CAT, linearized pACCEPT-F, restriction enzymes *SgfI* and *PmeI*, Promega Buffer C with BSA, T4 DNA ligase, ATP, and DTT. The restriction digest was initiated by incubating the reaction tube at 37°C for 1 hour. Following digestion, the reaction temperature was lowered to 25°C for 1 hour to allow for the ligation reaction to occur. Following ligation, an aliquot of the reaction was transformed into *E. coli* JM109 bacterial cells, and the transformation mixture was plated onto LB plates with ampicillin. Of the resultant colonies, 100 were re-plated onto LB plates with chloramphenicol. Colonies which grew on chloramphenicol contained the CAT gene. Transfer efficiency of the CAT gene from the pENTRY-4 CAT to the acceptor vector was determined to be approximately 37%.

Example III

An inducible system useful for cloning including directional cloning includes a recombinant host cell encoding a gene product regulated by an inducible promoter, which gene product specifically increases transcription of a DNA of interest in a vector introduced to the cell. In one embodiment, a first vector includes the open reading frame for a gene of interest operably linked to a promoter, e.g., a T7 promoter, which vector has a transcription terminator sequence, for instance, the *rrnB* terminator (to reduce aberrant expression), 5' to the promoter, a drug resistance gene, e.g., *kan*^R, sequences which permit the vector to be maintained in a host cell at high copy numbers, optionally sequences which reduce vector multimerization, e.g., *cer* sequences, as well as restriction enzyme sites flanking the open reading frame. In one embodiment, the restriction enzyme sites flanking the open reading frame are for two different infrequent cutters which do not generate complementary DNA ends (enzyme I and enzyme II) (Figure 21). The vector in Figure 21 also includes a T7 transcription terminator 3' of a *Pme*I site. A second vector having a backbone of interest for the open reading frame, preferably contains a different drug resistance gene, e.g., *amp*^R, and optionally the same transcription terminator sequences, promoter, sequences which permit the vector to be maintained in a host cell at high copy numbers, and optionally sequences which reduce vector multimerization as the vector containing the open reading frame of interest, wherein the transcription terminator sequences and promoter in the second vector are 5' to restriction enzyme sites for two restriction enzymes (enzyme III and enzyme IV) that generate ends that are compatible with ends generated by enzyme I and enzyme II, respectively. For instance, enzyme I is *Sgf*I, enzyme II is *Pme*I, enzyme III is *Pvu*I, and enzyme IV is *Dra*I. In another embodiment, the restriction sites recognized by enzymes I and III are the same, e.g., sites for *Sgf*I, and the restriction sites recognized by enzymes II and IV are the same, e.g., sites for *Pme*I. The resulting vector is introduced into a host cell which can be induced to express a gene product which increases transcription of the promoter which is 5' to the open reading frame, e.g., a gene product such as T7 RNA polymerase.

For example, a rhamnose-inducible system including a host cell useful to a clone and express a gene of interest was prepared. For instance, one or more of the *rhaBAD* catalytic genes in JM109 are deleted, replaced or interrupted via insertional mutagenesis. In one embodiment, the *rhaB* gene in JM109 was deleted, and a
5 vector with the *rhaBAD* promoter (e.g., see Egan et al., J. Mol. Biol., 234:87 (1993) and Wilms et al., Biotech Bioeng., 73:95 (2001)) linked to the T7 RNA polymerase open reading frame, stably introduced to those cells, yielding recombinant host cell JM109RX. A vector containing a luciferase gene linked to the T7 promoter was introduced to JM109RX, BL21(DE3) (Novagen), and BL21-AI (Invitrogen) cells.
10 The transformed cells were grown at either 25°C or 37°C, then contacted with rhamnose (JM109RX), IPTG (BL21(DE3)), or arabinose (BL21-AI), and luciferase activity measured at various time points.

The data showed that there was a much lower level of uninduced luciferase expression in transformed JM109RX cells than in the comparable arabinose
15 inducible system. The rhamnose inducible system may thus be particularly useful to clone toxic genes present in a donor vector or an amplified fragment, although the rhamnose-inducible system is not limited to the cloning of those genes.

Moreover, the induction of luciferase activity in transformed JM109RX cells was slow compared to luciferase activity in transformed BL21(DE3) or BL21-AI
20 cells, yet resulted in high final induction levels, e.g., high protein levels, e.g., at times $t = 4$ hours at which RLU were 100 X greater (Figures 22A and C). Further, the use of a rhamnose-inducible system at 25°C yielded more luciferase activity than at 37°C, e.g., at least 10-70 fold more at peak (Figures 22A and C). The observed expression profile of such a system may allow for increased solubility of
25 the expressed protein, e.g., due to increased time to fold. In addition, the rhamnose-inducible system is glucose repressible. Therefore, combinations of rhamnose and glucose may be employed to finely tune the expression profile of a gene of interest which is linked to a *rhaBAD* promoter.

30

Example IV

A system to express a toxic gene was prepared. A stably transformed host cell, JM109, was prepared that contained an expression vector encoding an immunity factor for barnase, barstar, which was expressed from a constitutive promoter, e.g., the 4c promoter, integrated into *lamB*. A vector containing a lambda
5 P_L promoter linked to a truncated barnase gene (see, e.g., Accession No. X12871 or M14442 (barnase genes from *Bacillus amyloliquefaciens*) or AE007600 (a barnase gene from *Alostridium acetobutylicum*), which lacked the secretory sequence, was introduced to those stably transformed cells.

10 All publications, patents and patent applications are incorporated herein by reference. While in the foregoing specification this invention has been described in relation to certain preferred embodiments thereof, and many details have been set forth for purposes of illustration, it will be apparent to those skilled in the art that the invention is susceptible to additional embodiments and that certain of the details
15 described herein may be varied considerably without departing from the basic principles of the invention.